2017

A Copula-based approach to differential gene expression analysis

Linda Akoth Chaba
Strathmore Institute of Mathematical Sciences (SIMs)
Strathmore University

A Copula-based Approach to Differential Gene Expression Analysis

Linda Akoth Chaba

Submitted in total fulfillment of the requirements for the Degree of Doctor of Philosophy in Biostatistics at Strathmore University

Strathmore Institute of Mathematical Sciences (SIMS)
Strathmore University
Nairobi, Kenya

June, 2017

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Approval

The thesis of Linda Akoth Chaba was reviewed and approved by the following:

Prof. Bernard Omolo.
Principal supervisor,
Division of Mathematics and Computer Science,
University of South Carolina Upstate.

Prof. John Odhiambo.
Co-supervisor,
Strathmore Institute of Mathematical Sciences,
Strathmore University.

Ferdinand Othieno.
Dean, Strathmore Institute of Mathematical Sciences,
Strathmore University.

Prof. Ruth Kiraka.
Dean, School of Graduate Studies,
Strathmore University.

Abstract

Microarray technology has revolutionized genomic studies by enabling the study of differential expression of thousands of genes simultaneously. The main objective in microarray experiments is to identify a panel of genes that are associated with a disease outcome or trait. In this thesis, we develop and evaluate a semi-parametric copula-based algorithm for gene selection that does not depend on the distributions of the covariates, except that their marginal distributions are continuous. A comparison of the developed method with the existing methods is done based on power to identify differentially expressed genes (DEGs) and control of Type I error rate via a simulation study. Simulations indicate that the copula-based model has a reasonable power in selecting differentially expressed gene and has a good control of Type I error rate. These results are validated in a publicly-available melanoma dataset. The copula-based approach turns out to be useful in finding genes that are clinically important. Relaxing parametric assumptions on microarray data may yield procedures that have good power for differential gene expression analysis.
Keywords: Copula; False discovery rate; Melanoma; Microarray; Power

Contents

List of Tables

List of Figures

Acknowledgment

First and foremost I would like to express my sincere appreciation to Prof. Bernard Omolo, for sharing his ideas so willingly and for being so dedicated to his role as my principal supervisor. My Ph.D. journey has been an amazing experience because he not only provided academic support but also opened my eyes to the world of genomic research. With his support and encouragement, I attended my first international conference in my first year of Ph.D. study. Not so often do Ph.D. students get the opportunity to attend international conferences in their first year of study.

Similar gratitude goes to Prof. John Odhiambo, my co-supervisor. Despite his busy schedule as the Vice Chancellor, he always had time for his students. I am thankful for his constant faith in my work. I am also grateful for his moral support when I felt like giving up.

I would like to acknowledge the Mwalimu Nyerere African Union scholarship Scheme (MNAUSS) for their funding support. I would also like to acknowledge the Strathmore Institute of Mathematical Sciences (SIMS) for offering me the opportunity to undertake doctoral studies and work at the same time. Without the support from the SIMS, I would not have made it this far. My thanks also go to my colleagues for sharing their experiences and providing moral support. I am indebted to Dr. William K. Kaufmann for the continuous and survival outcome data used in this study.

Finally, I am grateful to my parents, sisters, and brothers for their love and support, and for always believing in my ability to succeed. I would also like to thank all of my friends who supported and encouraged me to strive towards my goal. Special thanks go to my husband Dan and my son Jean. Without their patience and support, none of this would be possible.

Abbreviations

CIN	Chromosomal Instability Index
CMLE	Canonical Maximum Likelihood Estimation
DAVID	Database For Annotation And Integrated Discovery
DEG	Differentially Expressed Genes
DNA	Deoxyribonucleic Acid
EB	Empirical Bayes
FDR	False Discovery Rate
FWER	Family Wise Error Rate
MMM	Mixture Model Method
NHM	Normal Human Melanocytes
LIMMA	Linear Model For Microarray
SAM	Significance Analysis of Microarrays
SNP	Single Nucleotide Polymorphism
RNA	Ribonucleic Acid
PCR	Polymerase Chain Reaction

Chapter 1

Introduction

1.1 Background of Microarray Technology

Microarray technology has revolutionized genomic studies by enabling the study of differential expression of thousands of genes simultaneously. The main objective in microarray experiments is to identify a panel of genes that are associated with a disease outcome or trait. Microarray technology has increasingly gained application in biological and medical research, where their main application is in the classification of cells (tumor or normal cell). This section describes the biology of gene expression and how gene expression data is produced and processed. We also provide an extensive review of the existing statistical approaches available for the analysis of microarray data.

1.1.1 Biology of Gene Expression

Cells are the fundamental working units of every living organism. Their growth and division are controlled by the activity of the deoxyribonucleic acid (DNA). A DNA molecule is a double-stranded polymer composed of four basic molecular units called *nucleotides*. DNA from all organisms is made up of the same chemical and physical components. A DNA sequence is a particular arrangement of the base pairs in the DNA strand. The entire DNA sequence that codes for a living thing is called its *genome*. The genome does not function as one long sequence, but is divided into a set of genes. A *gene* is a segment of DNA that directs the synthesis of a protein. The expression of the genetic information stored in the DNA molecule occurs in two stages: (i) *transcription*, where the DNA molecule is transcribed into a ribonucleic acid (RNA); and (ii) *translation*, where RNA is translated into proteins that perform various cellular functions. These two processes describe the central dogma of biology which states that DNA makes RNA and RNA makes

protein (See Figure 1.1). The process of transcribing a gene's DNA sequence into RNA is called *gene expression*. A gene's expression level indicates the approximate number of copies of that gene's RNA that is produced in a cell. The measurements of the expression levels have been made easy with the introduction of microarray experiments.

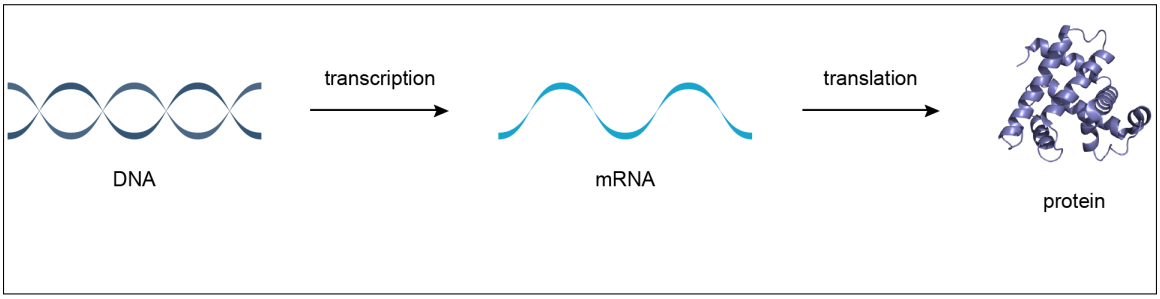

Figure 1.1: The Central Dogma of Molecular Biology: DNA makes RNA makes proteins. Image downloaded from `http://www.atdbio.com/content/14/Transcription-Translation-and-Replication\figure-central-dogma`

1.1.2 Microarray Experiment and Expression Data Generation

There are two types of microarray experiments: cDNA and oligonucleotide microarrays. The main difference between the two types of microarrays is in the components or molecules of DNA that are involved in the hybridization process. For cDNA microarrays, both the targets and probes are the cDNA molecules, while for the oligonucleotide arrays the targets are cDNA molecules and the probes are well-chosen small segments of cDNA, known as *oligos*. A sketch of the cDNA microarray technology is provided in Figure 1.2. Here, selected probes are amplified through the polymerase chain reaction (PCR) and the PCR product is printed to a glass slide using a high-speed robot. The glass slides consist of thousands of spots. Then mRNA are experimentally extracted from normal and tumor cells or reference and test cells, respectively. These are converted to cDNA through reverse-transcription, amplified by PCR and labeled using two florescent dyes (Cynine 3 or Cy3 (green) and Cynine 5 or Cy5 (red)). Test (tumor) cells are usually labeled with Cy5. These are mixed in equal proportions and allowed to hybridize with cDNA spotted in the glass slide. Once the hybridization is completed, the slides are washed and scanned with a scanning laser microscope, which is able to measure the brightest of each florescent spot. Brightness reveals how much of a specific DNA fragment is present on the target. Each spot represents a gene. Grey spots denote genes that were expressed in neither type of cell, while colored spots identify genes that were expressed in one of the cells or both.

The intensity of the color of the spot discloses the relative expression of the gene in the two cells. Usually a measurement scale is provided to associate each color tone with a log-transformed ratio between the expression levels in the two cells. The resultant data matrix consists of gene expression levels (rows) and replicated experiments (columns).

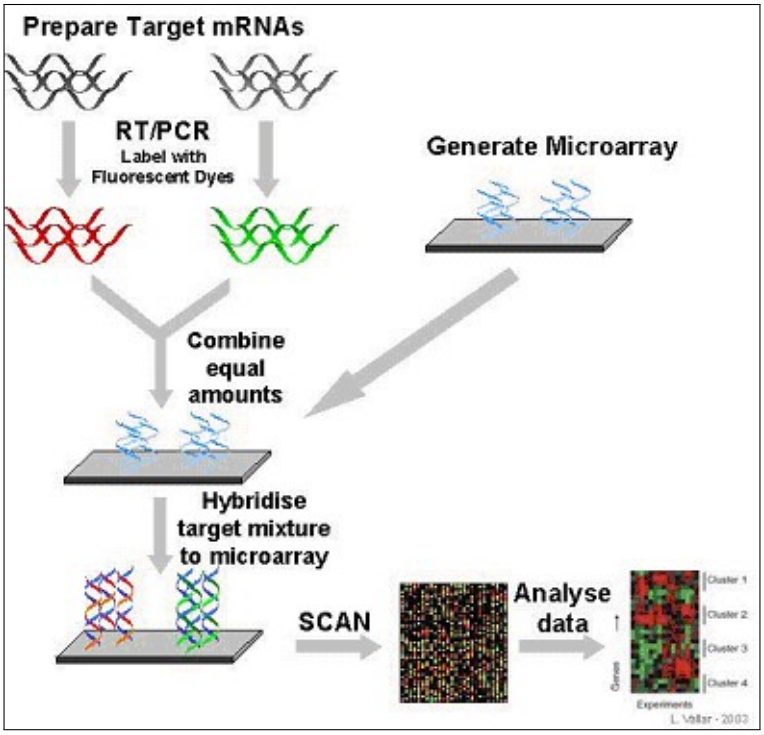

Figure 1.2: Overview of a cDNA microarray experiment. Image downloaded from `http://www.microarray.lu/en/MICROARRAY_Overview.shtml`

1.1.3 Preprocessing of Microarray Data

Data preprocessing in microarrays is done with the aim of reducing data variability and dimensionality (Sebastiani et al., 2003). Two processes are involved in preprocessing: normalization and filtering on either raw data or transformed data.

The goal of normalization is to remove systematic distortion across microarrays to render comparable the experiments conducted under different conditions. Normalization techniques can be used either locally or globally. Global normalization uses all genes in the microarray to identify a transformation of the expression while the local normalization uses only the house keeping genes (genes known to remain constantly expressed in different experimental conditions).

The goal of filtering is to reduce variability by removing genes that have measurements

that are not sufficiently accurate and to reduce dimensionality of the data by removing gene that are not sufficiently differentiated. The choice of genes to be removed can differ substantially according to the microarray platform and the technique chosen for analysis. For cDNA microarrays, genes with negative or small expression values are removed while for the Affymetrix platform, all genes labeled A (Absent) or M (Minimal) are removed.

Transformation of raw data is recommended since the corrected intensity values are highly skewed. It is assumed that log transformation produces normally distributed data (Nadon and Shoemaker, 2002). The best transformation method is still an open problem though.

1.1.4 Identification of Differentially Expressed Genes

A gene is differentially expressed if its expression level is associated with a response or a covariate of interest. The covariates could be the type of cell, the type of drug, etc., while responses could be survival time or any other clinical outcome in the context of other clinical studies.

Microarray data analysis methods can be subdivided into two broad categories: unsupervised and supervised methods. Unsupervised analysis or class discovery is an unbiased analysis of microarray data. No prior phenotype information is used and clustering methods are used to group the samples based solely on the microarray data. In a supervised analysis (also called class prediction), previous knowledge is taken into account. Its aim is to identify genes or develop a model that is able to assign patients to different classes based on the microarray data. Ding (2003) proposed and studied an unsupervised method to select relevant genes based on their similarity information only. The method relied on a mechanism for discarding irrelevant genes. When applied to expression profiles of colon cancer and leukemia, their unsupervised method selected relevant genes close to those selected using supervised methods. The existing literature on the application of discriminant and cluster analyses include Eisen et al. (1998) and Golub (1999). Eisen et al. (1998) used cluster analysis to find patterns of gene expression. However, they did not develop methods for modelling gene expression levels through a suitable statistical model. Even though they are limited in their applications, unsupervised methods are still used for pattern discovery and dimension reduction.

Supervised methods can further be classified as parametric, non-parametric and semi-parametric statistical methods. For any statistical method to be useful, it must have the following characteristics: ability to quantify the degree of association and the corresponding statistical significance between each gene expression level and the outcome of interest

(covariate); the ability to control the overall error rate; and robustness against outliers and model misspecification (Jung et al., 2005).

Among the early methods used was the fold-change method (DeRisi et al., 1996; Schena et al., 1996). This approach identifies genes as differentially expressed if the difference in expression levels between two conditions is greater than some specified threshold. This method has an advantage of being simple. However, it is known to be unreliable because statistical variation is not taken into account. It is also subject to bias if the data is not properly normalized (Sreekumar, 2008).

A number of parametric approaches exists in the literature. The most common and a straightforward approach is the traditional t-test for two samples. It compares the difference between two means in relation to the variation in the data. Its advantage over the fold-change method is that, it is easy to calculate p-values and confidence intervals (Bair, 2013). However, it has challenges when applied to data with small sample size. In particular, there is overestimation of variance when the sample size is small. Some of these parametric models have been assessed for goodness-of-fit to automatically detect outliers that possess too large deviation from the overall pattern (Li and Wong, 2001).

Given the challenges faced by t-test, a number of authors have proposed alternative methods to identify differentially expressed genes. Bayesian statistical methods have been developed for differential gene expression with a view to, *inter italia*, finding significant genes or gene signatures in large oncological microarray studies. These methods combine information across genes to avoid inaccurate variance estimates as a result of small sample sizes (Bair, 2013). Baldi and Long (2001) developed a Bayesian probabilistic framework for microarray data analysis. They modeled log-expression values by independent normal distributions, parameterized by corresponding means and variances with hierarchical prior distributions. Simulations showed that point estimates, combined with a t-test, provided a systematic inference approach that compared favorably with simple t-test or fold methods, and partly compensated for the lack of replication. Newton et al. (2001) described a version of parametric Empirical Bayes (EB) analysis for spotted microarrays and was restricted to the single-slide data in which each gene produces two measurements, one from each cell condition. Ibrahim et al. (2002) developed a Bayesian model for analyzing microarray data and used it to identify a subset of genes that are differentially expressed between normal and cancer cells. Lee et al. (2003) developed hierarchical Bayesian models for gene selection for binary data and applied the method to cancer classification via cDNA microarrays to identify significant genes. Extending Lee et al. (2003) work, Kendziorski et al. (2003) proposed a general EB modelling approach which allows for replicate expression profiles in multiple conditions. The hierarchical

mixture model accounts for differences among genes in their average expression levels, differential expression for a given gene among cell types, and measurement fluctuations. Smyth (2004) proposed a linear model for microarray (LIMMA) method that uses empirical Bayes test statistic. Scharpf et al. (2009) developed hierarchical Bayesian models for gene selection for binary data collected from different studies, and used it to identify significant genes. They applied their model to four breast cancer studies using different technologies (cDNA and Affymetrix) to estimate differential expression in estrogen receptor-positive tumors versus estrogen receptor-negative tumors. Results from their study showed a strong evidence that borrowing strength across both genes and studies can be effective in the analysis of multiplatform studies. Bayesian methods have proven to be very efficient in situations where the number of observations is small, as is the case for most microarray studies (Jeffery et al., 2006). However, Bayesian models do not distinguish genes with low levels of differential expression from those with no differential expression well (Scharpf et al., 2009).

A class of non-parametric methods has been proposed to identify DEGs. This is where the distribution of random errors are estimated without parametric assumptions. This idea has been proposed by a number of researchers. Tusher et al. (2001) used the statistical analysis of microarrays (SAM) method, which identified genes with statistically significant changes in expression. This method assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measures. For genes with scores greater than an adjustable threshold, SAM uses permutations of the repeated measurements to estimate the percentage of genes identified by chance, the false discovery rate (FDR). This method is capable of addressing problems with the fold-change approach but the estimation of variance can be affected if a small sample size is used. Efron et al. (2001) applied a non-parametric EB method in order to identify DEGs. Their method used a simple non-parametric mixture prior to model the population of genes affected by radiation or not, thereby avoiding parametric assumptions about gene expression. They found a close connection between the estimated posterior probabilities and a local version of the FDR, thereby allowing for the analyst to handle multiple testing issues that arise when dealing with a large number of simultaneous tests. Le et al. (2003) proposed a non-parametric statistical approach, called the mixture model method (MMM), to handle the problem when there are a small number of replicates under each experimental condition. They compared their method with SAM and showed that their method was better than the SAM approach. Pan (2003) proposed a non-parametric method to detect differential gene expression for replicated microarray experiments conducted under two conditions. Their method aimed at constructing a null

statistic that estimates a null distribution of a test statistic directly. They assessed their method with the existing methods when applied to the SAM, the MMM and the EB methods. In the process, they compared the performances of the three methods (SAM, MMM, EB) with each other. They showed that the SAM method was more robust to the use of the null statistic than the MMM method. They also showed that their method performed better than the existing methods in approximating the null distribution of the test statistic. However, their results were comparable to the existing methods when applied to the real dataset.

Semiparametric models have also been used to analyze gene expression data. Dhanasekaran et al. (2001) applied a method that identified a prognostic gene with a p-value calculated by fitting a Cox regression model without adjusting for multiplicity of the original genes. Wigle et al. (2002) used the Cox proportional hazards model to determine patterns of gene expression segregating with clinical outcome. They fitted a univariate Cox regression model on each gene expression level and adjusted p-values for the multiple testing procedure based on Dubey's approach (Dubey (1994)). However, Cox regression methods may not be robust in the presence of outliers (Owzar et al., 2007).

Newton et al. (2004) proposed a semiparametric hierarchical mixture method (HMM) to detect differentially expressed genes. Compared to several competing methodologies, their methodology exhibited good operating characteristics in a simulation study, on the analysis of spike-in data, and in a cross-validation calculation. Owzar et al. (2007) proposed a copula model for differential gene expression. This model assumes a parametric dependence between individual gene and time to event outcome. It was developed to identify individual gene expression associated with the time-to-event outcome. This method adequately controlled family wise error rate (FWER).

A number of authors have compared and evaluated the performances of different methods for gene expression analysis. These include (Dudoit et al., 2002; Troyanskaya et al., 2002; Schwender et al., 2003; Qin and Kerr, 2004; Jeffery et al., 2006; Kim et al., 2006; Sreekumar, 2008; Jeanmougin et al., 2010; Bair, 2013; Bandyopadhyay et al., 2014). Troyanskaya et al. (2002) addressed in their study the problem of robust identification of differentially expressed genes from a microarray data. They compared the performance of three non-parametric tests: non-parametric t-test, Wilcoxon rank-test and a heuristic method based on rank-sum test. They showed that all the methods exhibited low false positive rates but the rank-sum test proved to be the most conservative method. Jeffery et al. (2006) compared 10 methods to find differentially expressed genes on 9 different datasets. They reported that the classification success of the methods are influenced by the feature selection method, the number of genes in the genelist, the number of cases

(samples) and the noise in the dataset. Bair (2013) discussed a number of statistical methods, including fold change, methods based on t-test and Bayesian methods that can be used to find differentially expressed genes. However, he did not compare their performance on any dataset. Bandyopadhyay et al. (2014) compared a number of parametric and non-parametric tests based on simulated datasets. They concluded that the selection of genes depends on the choice of statistical technology and that the performance of the methods are affected by the samples size, number of replicates and distributional assumptions among other factors.

Identification of DEGs from combined microarray studies have also been common. Conlon et al. (2007) integrated information from different studies using a joint stochastic model for the available data. Scharpf et al. (2009) adopted Conlon et al. (2007) work and developed a hierarchical Bayesian model for gene selection for binary data collected from different studies, and used it to identify important genes expression. Their study assumed that the genes are independent. This approach of combining data from several independent microarray studies is termed as "microarray meta-analysis" (Tseng et al., 2012). Heterogeneity across studies is a major concern in carrying out microarray meta-analysis. Normalizing across studies and directly merging datasets for differentially gene analysis is one way of handing heterogeneity. This approach, however, restrict selection of studies from same or similar array platforms (Tseng et al., 2012). Other existing methods for carrying out a microarray meta-analysis include: combining p-values, combining effect sizes, combining ranks and combining latent variables.

1.2 Error Rate Control in Microarray Studies

A typical microarray experiment involves testing several hypotheses simultaneously. In this case, the probability of Type I error increases with the increase in the number of hypothesis to be tested. A global test of significance should therefore be conducted to determine if there is any significant value in the set of estimated parameters. Adjusting the p-value for the number of hypotheses implies controlling for Type I error.

Control of Type I error rate under multiple testing was initially done using family wise error rate(FWER) by Westfall and Young (1993). Among the first authors who adopted this method to identity differentially expressed gene was Dudoit et al. (2002). The FWER has been a useful approach in controlling Type I error, however, it is too conservative. This shortfall led to the development of an alternative method, the false discovery rate (FDR). FDR was developed by Benjamini and Hochberg (1995) as an improvement on the FWER approach. FDR is defined as the expected false discoveries among all the tests

that are called significant.

Table 1.1: Possible outcomes for testing G hypotheses for significance

	Null True (H_0)	Alternative True (H_1)	Total
Tests not Significant	U	T	$G-R$
Tests called Significant	V	S	R
Total	G_0	$G-G_0$	G

In Table 1.1, assume there are G hypotheses to be tested, R is the number of rejected hypothesis and G_0 is the number of true null hypotheses (an unknown parameter). $G-G_0$ is the number of true alternative hypotheses, V is the number of false positives (Type I error) (also called "false discoveries") and S is the number of true positives (also called "true discoveries"). T is the number of false negatives (Type II error), U is the number of true negatives and $R = V + S$ is the number of rejected null hypotheses (also called "discoveries", either true or false). The false discovery rate is defined as (Storey, 2002)

$$FDR = E\left[\frac{V}{R}|R > 0\right] \tag{1.1}$$

A number of authors have worked on FDR (Yekutieli and Benjamini, 1999; Efron et al., 2001; Storey, 2002; Storey and Tibshirani, 2003; Schwartzman et al., 2008; Schwartzman and Lin, 2011), among others.

For microarray studies, the positive false dicovery rate (pFDR) is preferred to other methods of controlling error. Storey (2002) introduced the q-value, which is the pFDR analogue of the p-values. They argued that the pFDR and the q-value were the most appropriate false discovery rate quantities to use . Storey and Tibshirani (2003) modified Benjamini and Horchberg's approach by estimating π_0. π_0 is the overall proportion of true null hypothesis in the study and $1-\pi_0$ is the proportion of significant results in the study. He showed that $\pi_0 = 1$ in Benjamini and Horchberg's approach.

The procedure below outline the steps followed in estimation of q-values given a list of p-value (Storey and Tibshirani, 2003).

1. Let $p_{(1)} \leq p_{(2)} \leq ...p_{(G)}$ be the ordered p- values. This also denotes the ordering of the features in terms of their evidence against the null hypothesis.

2. For a range of λ , say $\lambda = 0, 0.01, 0.02, ... , 0.95$, calculate

$$\hat{\pi}_0(\lambda) = \frac{\#(p_j > \lambda)}{G(1-\lambda)}.$$

3. Let $\hat{f}$ be the natural cubic spline with 3 df of $\hat{\pi}_0(\lambda)$ on λ

4. Set the estimate of π_0 to be $\hat{\pi}_0 = \hat{f}(1)$.

5. Calculate $\hat{q}(p_{(G)}) = \hat{\pi}_0 p_{(G)}$.

6. For $i = G-1, G-2, ...1$, calculate

$$\hat{q}(p_{(i)}) = \min\left\{\frac{\hat{\pi}_0.\mathrm{G}.p_{(i)}}{i}, \hat{q}(p_{(i+1)})\right\}.$$

7. The estimated q-value for the i^{th} most significant feature is $\hat{q}(p_{(i)})$.

1.3 Statement of the Problem

Despite all the proposed methods mentioned above, there is no unanimous agreement on any particular gene selection method. Some of the methods require normality assumptions, which may be violated in practice. Furthermore, most methods were developed for finding differentially expressed genes based on groups or class prediction based on discrete categories. However, there are some outcomes of interest that are continuous in nature.

There is need to develop more methods that take into account continuous outcomes and at the same time relax the normality assumptions. In this work, we propose a semi-parametric approach that does not rely on the normality assumption for the marginal distributions.

1.4 Objectives of the Study

The main aim of this study is to develop a copula model for variable selection in high-dimensional data (large p, small n) and apply it in differential gene expression analysis from microarray studies.

The specific aims of this study are to:

(i) Review and compare methods for differential gene expression analysis with regard to a quantitative outcome.

(ii) Develop and apply a copula model to obtain prognostic gene signatures that are associated with a quantitative trait.

(iii) Compare the developed model in Aim (ii) to standard parametric methods reviewed in Aim (i).

1.5 Thesis Outline

This thesis is organised in chapters. The content of the remaining chapters are summarized as follow:

In **Chapter 2**, a review and evaluation of methods for finding differentially expressed genes in the presence of a quantitative outcome is performed. Four methods are discussed and applied on both simulated and a real dataset to asses their performance.

In **Chapter 3**, we introduce a copula model. Families of copulas are also described as well as the available methods of choosing the "best" copula in terms of goodness-of-fit. Some areas of applications are also discussed briefly. In **Chapter 4**, we develop a copula-based algorithm for finding differentially expressed genes. Simulated datasets are used to asses the power of the developed copula approach. We then apply it to a melanoma dataset for validation.

In **Chapter 5**, the copula-based approach developed in chapter 4 is compared with the quantitative trait analysis (QTA) method for finding differentially expressed genes based on power and control of Type I error rate using simulated datasets. A summary of the findings, limitations and possible extensions are provided in **Chapter 6**.

Chapter 2

Literature Review

2.1 Background

Microarray technology has revolutionized genomic studies by enabling the study of differential expression of thousands of genes simultaneously. In the recent past, a number of statistical methods have been developed for class comparison and prediction, based on the gene expression profiling of tumors, cell-types, etc. One of the early methods developed was the fold-change method. This method did not account for statistical variation across the samples and suffered from bias if the data are not properly normalized (Sreekumar, 2008).

A number of articles have provided a survey of different statistical methods for finding differentially expressed genes (DEGs). See Chapter 1 Section 5.2 for more information. Despite all the surveys mentioned above, there is no unanimous agreement on any particular gene selection method as the standard. A review and comparison of the statistical methods may provide bioinformaticians and other biomedical researchers with a useful guide for choosing the right method for the right data in differential gene expression analysis. Furthermore, even though work has been done on the development of methods for the differential analysis of gene expression data measured in two conditions, open research questions still exist regarding the analysis of gene expression data in which the training signal is a continuous variable.

This chapter reports a comparative review of four methods: the SAM, the LIMMA, the lassoed principal components (LPC) and the quantitative trait analysis (QTA) and their comparison in identifying genes that are associated with a continuous outcome from the systems biology of melanoma, using a larger number of melanoma cell-lines than reported in Kaufmann et al. (2008). While the comparison of some of these methods has been done, most of them concentrated on finding gene signatures based on two groups. A comparison of the LPC method with other methods is conspicuously missing in almost

all the surveys presented in the literature. Furthermore, the available studies do not assess the biological and clinical significance of genes generated by theses methods. Our study attempts to fill this gap in the literature. The comparison is based on the size and the statistical assessment of the predictive and the prognostic properties of the genelists produced by these methods.

2.2 Review of Some Statistical Methods for Microarray Data

Most of the methods discussed were developed to identify genes that are expressed in varying biological conditions. In this section, we do an elaborate review of some of the commonly used methods that allow detection of differentially expressed genes with respect to a quantitative outcome.

2.2.1 Statistical Analysis of Microarrays (SAM)

The SAM method was originally developed to identify genes that are differentially expressed by incorporating a set of gene-specific t-tests. Although Tusher et al. (2001) analyzed a two-state experiment (with a dichotomous covariate or response), the SAM procedure can be applied to studies with continuous responses as well. The SAM method identifies DEGs by use of gene-specific moderated t-tests on the basis of the regression coefficient relative to the standard deviation of repeated expression measurements for that gene. SAM employs the false discovery rate (FDR) to control for the multiple testing problem and estimates the FDR through the permutation of values of the response variable and the moderated t-tests.

Let x_{ij} be the expression level for the i^{th} gene from the j^{th} sample and y_j be the covariate for the j^{th} sample. The linear model of analysis can be expressed as

$$x_{ij} = \beta_{i0} + \beta_{i1} y_j + \varepsilon_{ij}, \quad i = 1, 2, ..., G, \quad j = 1, 2, .., n. \tag{2.1}$$

Here, we assume that

$$\varepsilon_{ij} \sim N(0, \sigma_i^2), \quad i = 1, ..., G, \quad j = 1, ..., n. \tag{2.2}$$

Essentially, the procedure assigns a score, $d(i)$, to each gene, on the basis of the regression coefficient relative to the standard deviation of repeated expression measurements for that gene. The score, $d(i)$, is defined as

$$d(i) = \frac{b_i}{s_i + s_0} \tag{2.3}$$

where b_i and s_i are the estimates of β_{i1} and the standard error of b_i, respectively.

$$b_i = \frac{\sum_j y_j(x_{ij} - \bar{x}_i)}{\sum_j (y_j - \bar{y})^2}, \tag{2.4}$$

$$s_i = \frac{\hat{\sigma}_i}{\sqrt{\sum_j (y_j - \hat{y}_j)^2}}, \tag{2.5}$$

and

$$\hat{\sigma}_i = \sqrt{\frac{\sum_j (x_{ij} - \hat{x}_{ij})^2}{n-2}}, \tag{2.6}$$

where

$$\begin{aligned} \hat{x}_{ij} =& \hat{\beta}_{i0} + b_i y_j \\ \hat{\beta}_{i0} =& \bar{x}_i - b_i \bar{y}. \end{aligned} \tag{2.7}$$

s_0 is a small positive constant called the *fudge factor*, which is added to s_i in order to minimize the coefficient of variation. This is calculated as a quantile of the standard deviations, s_i. Efron et al. (2001) show that the optimum value of s_0 derived by cross-validation is the 90-th percentile of the distribution of the sample variance.

To find genes that are associated with the continuous outcome, the score $d(i)$ is calculated first from the original data and the values of $d(i)$ are ranked to obtain the observed order statistics $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(G)}$. The distribution of the $d_{(i)}$ under null hypothesis is unknown and is hence estimated by taking B permutations of the covariate and calculating the permuted expression scores $d^p_{(i)}$, $p = 1, ..., B$. The expected relative linear coefficient of regression, $\bar{d}_{(i)}$, is then obtained by averaging all $d^p_{(i)}$, that is $\bar{d}_{(i)} = \frac{1}{B}\sum_p^B d^p_{(i)}$. For a given threshold Δ, significant genes are identified as those for which $|d_{(i)} - \bar{d}_{(i)}| \geq \Delta$. Δ is chosen by cross-validation, as discussed in Tibshirani et al. (2003). Since some of the significant genes are identified by chance, an estimate of the expected rate of false positives (FDR) is needed. FDR is defined as

$$\text{FDR} = \text{E}\left(\frac{V}{R} \,\middle|\, R > 0\right) P(R > 0), \tag{2.8}$$

where V is the number of false positives and R is the number of genes declared significant (Benjamini and Hochberg, 1995). Let

$$d_0 = \max_{d_{(i)} \leq \bar{d}_{(i)} - \Delta} d(i) = \text{cut}_{\text{low}}(\Delta) \tag{2.9}$$

and

$$d_1 = \min_{d_{(i)} \geq \bar{d}_{(i)} + \Delta} d(i) = \text{cut}_{\text{up}}(\Delta). \tag{2.10}$$

Then the expected number of false positives obtained by chance is given by

$$V(\Delta) = \frac{\hat{\pi_0}}{B} \sum_{p=1}^{B} \left(I_{[d^p(i) \leq d_1]} + I_{[d^p(i) \geq d_0]} \right), \tag{2.11}$$

where $\hat{\pi_0}$ is an estimate of prior probability of no differential gene expression (Schwender et al., 2003). The FDR is then estimated as

$$\widehat{FDR} = \frac{V(\Delta)}{R(\Delta)}. \tag{2.12}$$

The SAM method is implemented in the *R* package called *samr*.

2.2.2 Linear Models for Microarray Analysis (LIMMA)

LIMMA is an *R* package that integrates a number of statistical methods to effectively analyse large gene expression data (Smyth, 2005). LIMMA fits a linear model for each gene, given a series of arrays, and uses the EB (Efron et al., 2001) method to estimate posterior variance for each gene (Smyth, 2004; Ritchie et al., 2015). The use of the EB method allows combination of information across genes thus improving variance estimation.

Let x_{ij} be the expression level for the i^{th} gene from the j^{th} sample and y_j be the covariate for the j^{th} sample. The linear model of analysis can be expressed as

$$x_{ij} = \beta_{i0} + \beta_{i1} y_j + \varepsilon_{ij}, \quad i = 1, 2, ..., G, \quad j = 1, 2, .., n. \tag{2.13}$$

Assume that

$$E(\mathbf{x}_i) = Y \boldsymbol{\beta}_i, \tag{2.14}$$

and

$$Var(\mathbf{x}_i) = W_i \sigma_i^2, \tag{2.15}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in})^T$, Y is a known design matrix and W_i is a known non-negative definite weight matrix. Certain contrasts are assumed to be of biological interest and are

defined as

$$\boldsymbol{\alpha}_i = C^T \boldsymbol{\beta}_i, \tag{2.16}$$

where C is a known contrast matrix. For each gene i, a linear model is fitted to obtain the coefficient estimator $\hat{\beta}_i$ and the variance estimator s_i^2. Estimated covariance matrices are given as $Var(\hat{\boldsymbol{\beta}}_i) = V_i s_i^2$, where V_i is a positive definite matrix not depending on s_i^2. The contrast estimator $\hat{\boldsymbol{\alpha}}_i$ is given as $\hat{\boldsymbol{\alpha}}_i = C^T \hat{\boldsymbol{\beta}}_i$ with estimated covariance matrices $Var(\hat{\boldsymbol{\alpha}}_i) = C^T V_i C s_i^2$. Prior information is assumed on σ_i^2 equivalent to a prior estimator s_0^2 with d_0 degrees of freedom, i.e.

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0 s_0^2} \chi_{d_0}^2. \tag{2.17}$$

This describes how the variances are expected to vary across genes (Smyth, 2004). Using Bayes' rule, the posterior variance becomes a combination of an estimate obtained from a prior distribution s_0^2 and pooled variance s_i^2,

$$\tilde{s}_i^2 = \frac{d_0 s_0^2 + d_i s_i^2}{d_0 + d_i}, \tag{2.18}$$

where d_0 and d_i are prior and empirical degrees of freedom. The posterior values shrink the observed variances towards the prior values with the degree of shrinkage depending on the relative sizes of the observed and prior degrees of freedom. The moderated t-statistic then becomes

$$\tilde{t}_{ij} = \frac{\beta_{ij}}{\tilde{s}_i \sqrt{v_{ij}}}, \tag{2.19}$$

where v_{ij} is the j^{th} diagonal element of $C^T V_i C$.

To asses the significance of each gene, the moderated t-statistics and their associated p-values are generally used (Ritchie et al., 2015). *limma* calculates the Bayesian log-odds of differential expression for each gene. The higher the value of the log-odds, the more significant the result. The family-wise error rate (FWER) and the FDR are used in multiple testing adjustment. The LIMMA method is implemented in the R package called *limma*.

2.2.3 Lassoed Principal Components (LPC)

The lassoed principal components (LPC) method involves using existing gene-specific scores (T) to calculate scores that provide a more accurate ranking of genes as differentially expressed (Witten and Tibshirani, 2008). Some of the gene-specific scores can be calculated using LIMMA (Smyth, 2005), SAM (Tusher et al., 2001) and standardized regression methods, among others existing methods. LPC identifies significant genes

based on the values of the FDRs. It estimates its FDR based on an adjustment of the FDR of the T (Witten and Tibshirani, 2008). The LPC method does not assume that genes are independent but rather takes into account that they work in pathways. The LPC method is similar to the LIMMA method in that they both combine information, or borrow strength, across genes. They do not also do permutation-based inference.

Let $\mathbf{X}$ be an $n \times G$ matrix of log-transformed gene expression levels, where n is the number of samples and G is the number of genes. Also, let $\mathbf{x}_i$ be the expression profile for gene i and $\mathbf{y}$ be the vector of quantitative outcomes. A gene-specific score, t_i, is calculated as the standardized regression coefficient of $\mathbf{y}$ onto $\mathbf{x}_i$. This is expressed as

$$t_i = \frac{\mathbf{x}_i^{'} y}{\sigma\sqrt{(\mathbf{x}'\mathbf{x})_{ii}}}. \tag{2.20}$$

A small constant σ is added to the denominator of the gene score in order to avoid a large ratio resulting from a small estimated standard deviation (Witten and Tibshirani, 2008).

To calculate the LPC scores, a model

$$t_i = \beta_0 + \sum_{j=1}^{n} v_i \beta_i \tag{2.21}$$

is fitted, where t_i is the gene-specific score calculated using standardised regression method as in (2.20), β_i is the multiple linear regression coefficient and v_i is the eige-narray of $\mathbf{x}_i$. The LPC score is the fitted value $\hat{t_i}$ obtained from model (2.21). The LPC approach can also be applied to studies with different outcome variables (e.g survival outcome, two-class or multiple-class type of outcomes).

The LPC algorithm is implemented in both the R package called *lpc* (Witten and Tibshirani, 2008) and BRB-ArrayTools (Simon et al., 2007).

2.2.4 Quantitative Trait Analysis (QTA)

This approach finds genes that are significantly correlated with a quantitative outcome such as age. It uses the Pearson's correlation or the Spearman's (rank) correlation coefficient as a measure of dependence to compute p-values.

Let X_{ij} be the expression level for the i^{th} gene from the j^{th} sample and y_j be the covariate for the j^{th} sample. The (linear) model of analysis can be expressed as

$$X_{ij} = \beta_{i0} + \beta_{i1} y_j + \varepsilon_{ij}, \quad i = 1, 2, ..., G, \quad j = 1, 2, .., n. \tag{2.22}$$

Here, we assume that

$$\varepsilon_{ij} \sim N(0, \sigma_i^2), \quad i = 1, ..., G, \quad j = 1, ..., n \tag{2.23}$$

and that the y_j values are fixed (not random). β_{i0} and β_{i1} represent the regression coefficients specific to gene i. For testing the significance of correlation for the i-th gene ($H_{0i} : \beta_{i1} = 0$ vs $H_{1i} : \beta_{i1} \neq 0$), we use the statistic T_i, defined as

$$T_i = \frac{\hat{\beta}_{i1}}{SE(\hat{\beta}_{i1})}, \tag{2.24}$$

where $SE(\hat{\beta}_{i1})$ is the standard error of $\hat{\beta}_{i1}$. T_i has the t-distribution with $n-2$ degrees of freedom. We reject H_{0i} if $\mid t_i \mid < t_{\alpha/2}$, $0 < \alpha < 1$. Equivalently, one can test for ρ_i, the correlation coefficient. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in})'$ and $\mathbf{y} = (y_1, y_2, ...y_n)'$. Given the observations (x_{ij}, y_j), the Pearson's correlation coefficient for the i^{th} gene, r_i, is defined as

$$r_i = \frac{n\sum_{j=1}^n x_{ij}y_j - \sum_{j=1}^n x_{ij}\sum_{j=1}^n y_j}{\sqrt{n\sum_{j=1}^n x_{ij}^2 - (\sum_{j=1}^n x_{ij})^2}\sqrt{n\sum_{j=1}^n y_j^2 - (\sum_{j=1}^n y_j)^2}}. \tag{2.25}$$

For testing $H_{0i} : \rho_i = 0$ vs. $H_{1i} : \rho_i \neq 0$, we use the statistic T_i^*,

$$T_i^* = \frac{r_i\sqrt{(n-2)}}{\sqrt{1-r_i^2}}, \tag{2.26}$$

which also has a t-distribution with $n-2$ degrees of freedom. Here r_i is the estimator of ρ_i. H_{0i} is rejected if $\mid t_i^* \mid < t_{\alpha/2}$, $0 < \alpha < 1$. With a simple algebraic manipulation, it can be shown that (2.24) and (2.26) are equivalent and so the latter was employed in this study.

There are two ways of controlling the number of false discoveries in the QTA approach. The first one is based on the p-values computed from the parametric $t-$ or $F-$tests. Here, a stringent p-value threshold (say $p < 0.001$), is used in controlling the number of false discoveries. The second approach uses multivariate permutation tests (Korn et al., 2004). The multivariate permutation tests are based on permutations of the covariate. For each permutation, the parametric test statistics are re-computed to determine a p-value for each gene. The genes are ordered by their p-values computed for each permutation, with genes having the smallest p-values appearing at the top of the list. For a pre-selected p-value threshold, the distribution of the number of genes that would have p-values smaller than that threshold is computed. That is the distribution of the number of false discoveries, since genes that are significant for random permutations are false discoveries. The algorithm selects a threshold p-value so that the number of false

discoveries is not greater than that specified by the user C percent ($C\%$) of the time, where C denotes the desired confidence level (Simon et al., 2007).

The QTA approach estimates false discovery rate (FDR) using the Benjamini and Horchberg's approach Benjamini and Hochberg (1995). For the i^{th} gene, the estimated FDR is given by

$$\widehat{FDR_i} = \frac{G \times p_i}{i}, \tag{2.27}$$

where p_i is the univariate p-value for the i-th most significant gene and G is the number of gene tested.

A concise summary of the four statistical methods is provided in Table 2.1.

Table 2.1: *Comparison of the four methods for finding DEGs. PB = permutation-based; NPB = non-permutation based.*

Method	**Test statistic**	**Error rate control**	**Inference**	**Statistical test**	**Inter-gene assumption**
SAM	Moderated-t	FDR	PB	Non-parametric	Independent
LIMMA	EB moderated-t	FDR/FWER	PB	Non-parametric	Dependent or Independent
LPC	LPC	FDR	NPB	Non-parametric	Dependent
QTA	t or F	FDR	Both PB and NPB	Parametric	Independent

2.3 Simulated Gene Expression Data

We conducted a simple simulation study to compare the four methods in terms of power. Let n and G denote the number of samples and genes, respectively. Further, let D denote the number of genes assumed to be truly differentially expressed. Then $(G - D)$ genes are assumed to be non-differentially expressed. The gene expression data matrix, $\mathbf{X}$, is a $G \times n$ matrix of log2-ratios. We can write $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$, where $\mathbf{X_1}$ and $\mathbf{X_2}$ are $D \times n$ and $(G - D) \times n$ matrices, respectively. We set $D = 50$, and $n = 35$ and G to be 1000. We generated the $(1000 - D)$ genes from the standard normal distribution. To generate the D genes, we used the standard normal distribution in conjunction with the Cholesky decomposition Golub and Van Loan (1996) of their correlation matrix as follows:

1. We generate an unstructured correlation matrix $\mathbf{\Omega}$. $\mathbf{\Omega}$ is a $(D+1) \times (D+1)$ matrix that has $(i, j)^{th}$ element given by $\omega_{i,j} = \text{corr}(x_i, x_j)$
2. Find the Cholesky factor, $\mathbf{A}$, of $\mathbf{\Omega}$ such that $\mathbf{\Omega} = \mathbf{A}\mathbf{A}'$.
3. Let $\mathbf{z}_i \sim N(\underline{0}, \mathrm{I}_n), i = 1, 2, ..., (D + 1)$.
4. $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{D+1})'$
5. $\mathbf{X}_{D+1} = \mathbf{A}\mathbf{Z}$.

$\mathbf{X}_{D+1}$ is the gene expression matrix for D genes that are assumed to be differentially expressed or significantly correlated with the covariate $\mathbf{y}$. $\mathbf{y}$ can take any of the $D + 1$ row vectors from the matrix $\mathbf{X}_{D+1}$. $\mathbf{X}_1$ is therefore a submatrix of $\mathbf{X}_{D+1}$ with dimensions $D \times n$. This simulation set-up assumes that each gene is observed across each sample. An R code for the above simulation is available in the Appendix B.

All the four methods are applied to the simulated data. Differentially expressed genes are identified based on the methods' estimated FDR values. A gene is differentially expressed if its estimated FDR is less than a pre-specified value α. Power is calculated as the ratio of the number of correctly identified differentially expressed genes, true positives (TP), to the total number of truly differentially expressed genes (Owzar et al., 2007).

2.4 Application

An analysis on a real microarray dataset is performed to evaluate how the methods perform in a real situation. The four methods are applied to the melanoma cell lines

dataset to identify DEGs. The genelists generated by the four methods are then applied to an independent melanoma dataset for prognostic assessments. Below are the descriptions of the datasets used in the application.

2.4.1 Data

Melanoma Cell Lines Dataset

The gene expression data (raw intensities) consists of 54 cell-lines (35 melanoma cell lines and 19 normal human melanocytes (NHMs)), each with 45,015 probes. This data is publicly available from Gene Expression Omnibus (`http://www.ncbi.nlm.nih.gov/geo/`) under accession number GSE40047. Only the melanoma cell lines are analyzed. The raw dataset is median-normalized and $\log_2$ transformed. If multiple probes map to the same gene symbol, they are reduced to one per gene symbol by using the most variable probe(set) measured by interquartile range (IQR) across arrays. Filtration and normalization of the gene expression data is implemented using BRB-ArrayTools software (Simon et al., 2007). A gene is filtered out if less than 20% of its expression data values has at least 1.5-fold change in either direction from the genes median value. Genes with more than 50 % missing data across all its samples are also filtered out. There are 3,860 genes available for subsequent analysis.

G_2 Checkpoint Function

Having obtained the gene expression data, we need to quantify the biological process in melanoma progression. We select the G_2 checkpoint function in this regard. The G_2 checkpoint is a position of control in the cell cycle that delays or arrests mitosis when DNA damage by radiation is detected. The G_2 checkpoint prevents cells with damaged DNA cell from entering mitosis, thereby providing the opportunity for repair and stopping the proliferation of damaged cells. Figure 4.1 below shows the four phases of the cell cycle, including the location of the G_2 checkpoint as the last checkpoint before mitosis. The G_2 checkpoint function scores were obtained from Kaufmann's lab (UNC - Pathology and Lab Medicine) and had been calculated as ratios of mitotic cells in 1.5 Gy ironizing radiation (IR)-treated cultures in comparison to their sham-treated control (i.e. IR to sham ratio) (Kaufmann et al., 2008). It had been shown in Omolo et al.Omolo et al. (2013) that the G_2 gene signature was prognostic for the development of distant metastasis, hence the choice of G_2 checkpoint function for this study.

Independent Melanoma Dataset

An independent dataset, consisting of gene expression data from 6307 genes on 58

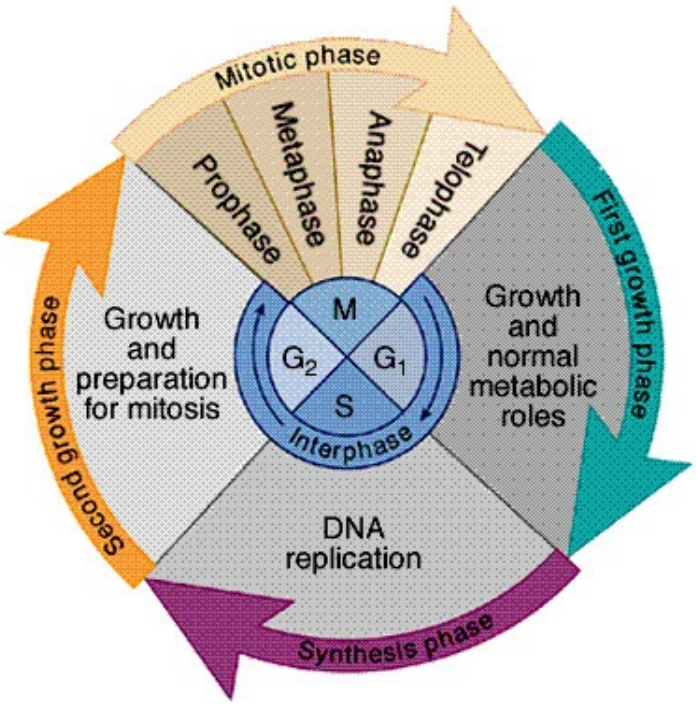

Figure 2.1: **Cell cycle**. After completing DNA synthesis and progression through the G_2 phase, the cell enters the mitotic phase, where the chromosomes segregate into two daughter cells. Image downloaded from `http://www.bristol.k12.ct.us/page.cfm?p=7093`.

primary melanomas with survival outcome, is obtained for assessing prognosis of the gene signatures from the four methods. This data set has been reported in Winnepenninckx et al. (2006) and will hereafter be referred to as the Winnx dataset. This data is publicly available in the Array Express data repository at the European Bioinformatics Institute (http://www.ebi.ac.uk/arrayexpress/) under the accession numbers: E-TABM-1 IGR_MELANOMA_STUDY. The primary endpoint for the study was a 4-year distant metastasis-free survival (DMFS), which was defined as the time interval between the diagnosis of the primary cutaneous melanoma and a distant metastasis or death from melanoma within 4 years. Patients alive at the date of last follow-up were censored at that date. Patients were also separated into two groups, one group with distant metastasis-free survival of more than 4 years (group M-) and one group with distant metastasis-free survival of 4 years or less (group M+).

2.4.2 List of DEGs

To find DEGs, we apply different software for different methods. For the LIMMA approach, we use *limma R* package. We fix the degrees of freedom for the design matrix to be 5. For the SAM approach, the *samr R* package is used. Δ is fixed at 0.00, to allow a large list of DEGs to be generated at different estimated FDR values. The number of nearest neighbors to use for imputation of missing features (*knn.neighbors*) is set at 10 and the number of permutations is fixed at 1000. The QTA method assesses significance of correlation based on Spearman's correlations and implements the procedure using the BRB-ArrayTools software. Similarly, the LPC method is implemented by the BRB-ArrayTools software. The number of DEGs are generated at various levels of estimated FDR threshold (0.01, 0.05 ,0.1, 0.2) for all the methods.

2.4.3 Prediction and Prognosis

We assess the predictive quality of each of the genelists by its mean squared error (MSE) of prediction of the G_2 checkpoint function. For this, linear models containing significant genes are formulated. Since $G >> n$, the least absolute shrinkage and selection operator (LASSO) algorithm (Tibshirani, 1996) is used to select genes to include in the models. LASSO builds a sequence of models containing upto n genes and index by F, the number of algorithmic steps relative to the model containing n genes (full model). For each F, a cross-validation estimate is obtained using the leave-one-out cross-validation (LOOCV) method. The final model selected corresponds to the F-value with the minimal estimated mean squared error.

We perform a survival risk prediction (SRP) to assess the clinical significance of the genelists using the Winnx dataset. The clinical outcome for this dataset was 4-year distant metastasis-free survival (DMFS) and the objective was to predict a patient's risk (low/high) for developing distant metastasis within 4 years of primary diagnosis. The SRP procedure entails first reducing the number of candidate genes to only the Cox ones, using the supervised principal component (SPC) method of Bair and Tibshirani (2004). These Cox genes are then used to compute the prognostic index for each sample. Samples (patients) with a prognostic index above the median are classified as high risk; otherwise, they are low risk. A log-rank test is performed to test if the two survival curves for the low- and the high-risk groups are significantly different, using the original DMFS values. A genelist would be prognostic for DMFS if the log-rank test is significant. The entire SRP procedure is implemented by a tool of the same name in BRB-ArrayTools software (Simon et al., 2007). We compare the performance of the genelists produced by the four

methods in survival risk prediction for the 58 samples in the Winnx dataset.

In addition, we use the Prediction Analysis of Microarrays (PAM) tool to predict the group membership of the 58 samples. Samples were grouped into two classes: a group with distant metastasis-free survival of more than 4 years (group M-) and a group with distant metastasis-free survival of 4 years or less (group M+). PAM uses the shrunken centroid algorithm developed by Tibshirani et al. (2002). This algorithm builds a number of linear models and selects the model with the least prediction error. A cross-validation estimate is obtained by using leave-one-out cross-validation (LOOCV). The entire model building process is repeated for each leave-one-out training set. The misclassification rate is calculated as the proportion of times the models incorrectly predict the class of the excluded samples. The genelist with the lowest misclassification rate is considered a good list for predicting a sample as belonging to group M+ or M-.

2.5 Results and Discussion

2.5.1 Differentially Expressed Genes

Each of the four methods is applied to the simulated data. The total number of genes that are correctly identified as differentially expressed, true positives (TP), are recorded at different estimated FDR levels. With the known number of TP, the power is also calculated to aid in comparison. Table 2.2 shows the number of DEGs and power by different methods at different FDR levels. The LPC method turns out to be the least powerful of all the methods. The SAM and the QTA methods are the most powerful methods in the identification of DEGs. The LIMMA method has moderate power (> 0.7) for the FDR thresholds considered, except at the FDR < 0.01.

Although the SAM and the QTA methods performed the best with the simulated dataset, we need to determine how they behave with a real dataset. We apply the methods to the melanoma cell lines dataset (in 2.4.1). The results are different from the ones obtained using the simulated dataset. We observe that while the QTA method did well with the simulated dataset, its performance is the worst in the identification of DEGs using the real dataset. In terms of power, the SAM method is still the best followed by the LIMMA method.

The difference in the performance of the QTA method when applied to the simulated and the real datasets could be explained by the fact that the simulated dataset is generated from a standard normal distribution. The QTA method strongly assumes that the gene expression levels (log2-ratios) are normally distributed. Gene expression data may violate this assumption. The LPC method assumes that a large set of genes work

Table 2.2: *Number of DEGs generated by the SAM, LIMMA, LPC and QTA methods at different levels of estimated FDR.*

	Simulated dataset				Melanoma dataset			
Estimated FDR (α)	SAM	LIMMA	LPC	QTA	SAM	LIMMA	LPC	QTA
0.01	49 (0.98)	6 (0.12)	0 (0.00)	50(1.00)	0	8	3	0
0.05	52 (1.00)	39 (0.74)	0 (0.00)	51(1.00)	33	16	4	0
0.1	53 (1.00)	50 (0.88)	0 (0.00)	54 (1.00)	33	22	7	4
0.2	56 (1.00)	57 (0.82)	0 (0.00)	67 (1.00)	173	55	24	56

together in a pathway to cause an outcome. In cases where this assumption is not met i.e. when only one gene or very few genes causes the outcomes, the LPC method loses power in selecting significant genes. This could explain the low performance of the LPC method in both simulated and real datasets. One disadvantage of the LPC method is that it does not rank genes using a metric that is relevant or truly of interest. It rather finds genes that generate high values when standard scores are projected into a high-variance subspace of the gene expression data (Witten and Tibshirani, 2008).

Since different spots on the microarrays are assumed to contain different probes (in the case of cDNA arrays) or different oligos (in the case of high-density oligonucleotide arrays), the expression of genes are assumed independent on these spots, even though some probes may represent the same gene and have dependent expression profiles. Consequently, not all the methods for selecting DEGs assume that the genes are independent. In particular, the LPC method does not assume that the genes are independent, while the SAM and the QTA methods do assume independence. While the LIMMA approach assumes independence, it works well when the genes are assumed dependent as well (Smyth, 2004). This has been one of the main differences among the four methods.

Before analyzing the validation datasets, the gene expression data were filtered and normalized to eliminate genes that were not sufficiently differentially across the samples and to correct for sample-specific bias (due to experimental artefacts/errors) and render the samples comparable, respectively. After normalization, the resulting expression data was log2-transformed so as to achieve a symmetric error distribution. Figure 2.2 shows the error distribution for four randomly selected melanoma cell lines and primary tumors as symmetric and can be regarded as "approximately" normal.

Figure 2.3 shows the number of overlapping genes from the four methods. It is very common to find a very low number of overlapping DEGs between multiple methods (Jeffery et al., 2006; Andrew et al., 2015).

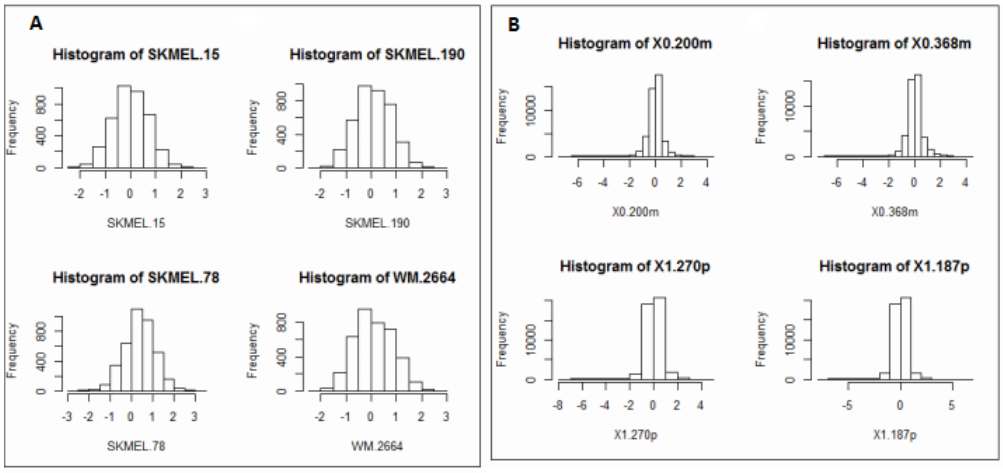

Figure 2.2: Error distribution for four of the melanoma cell lines (A) and primary tumors (B). The histograms are fairly symmetric and would approximate the normal distributions, considering the number of genes in each dataset.

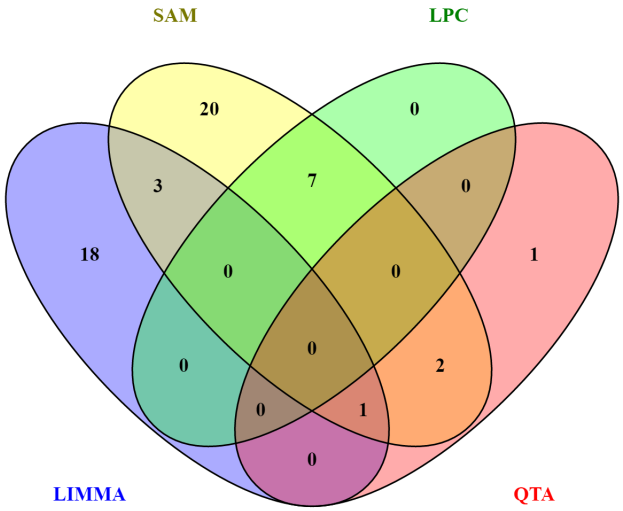

Figure 2.3: Number of overlapping genes from the SAM, the LIMMA, the LPC and the QTA genelists based on the melanoma cell lines dataset.

2.5.2 Prediction and Prognosis

We use the genelists generated by the four methods to build linear predictive models for the G_2 checkpoint function, via the LASSO with LOOCV. Table 2.3 is a summary of the results. The QTA genelist turns out to be the best in predicting G_2 followed by

the SAM genelist, then the LIMMA genelist. In order to get additional insight into the performance of the four methods, the four genelists are combined to get 52 unique genes. This combined genelist yields an r^2 of 0.5. A combination of all the genelists has a much better performance than most of the genelists generated by the individual methods.

Table 2.3: *Comparison of G_2 checkpoint function prediction by the SAM, LIMMA, LPC and QTA genelists generated at $\alpha = 0.1$. The number of genes associated with DMFS (Cox genes) are also included.*

Method	**# Genes in model**	r	p	R^2	**Adjusted R^2**	**# Cox genes**
SAM	10	0.652	<0.001	0.43	0.193	5
LIMMA	6	0.550	0.0006	0.3	0.150	1
LPC	3	0.421	0.0117	0.18	0.100	1
QTA	4	0.721	<0.001	0.52	0.456	1
Combine	16	0.710	<0.001	0.5	0.105	6

Gene expression data for the four genelists are extracted from the Winnx dataset for performing survival risk prediction. The difference between the survival curves for the low- and high-risk groups is significant for the SAM genelist (log-rank $\chi^2 = 5.5, P = 0.019$), the LPC genelist (log-rank $\chi^2 = 5.7, P = 0.0166$) and the QTA genelist (log-rank $\chi^2 = 4.8, P = 0.0374$) but not for the LIMMA genelist (log-rank $\chi^2 = 0.1, P = 0.791$). Results are shown in Figure 2.4.

We further subjected the combined genelist to a survival risk prediction analysis using the Winnix dataset. This genelist provides a good prediction of the G_2 checkpoint function and is the most prognostic genelist (log-rank $\chi^2 = 8.5$, P= 0.00351, **Fig 2.5**). We also observe that the misclassification rates based on PAM analysis are high for all the genelists. The misclassification rates are as follows: 36%, 41%, 31% and 36% for the SAM, LIMMA, LPC and QTA methods respectively (**Tab 2.4**).

Table 2.4: *Misclassification rates based on the Prediction Analysis for Microarrays (PAM).*

	SAM	LIMMA	LPC	QTA
Misclassificarion rate	36%	41%	31%	36%

2.6 Conclusion

In this chapter, we compare four methods (SAM, LIMMA, LPC and QTA) for identifying DEGs in terms of their power to detect differential gene expression, the predictive ability

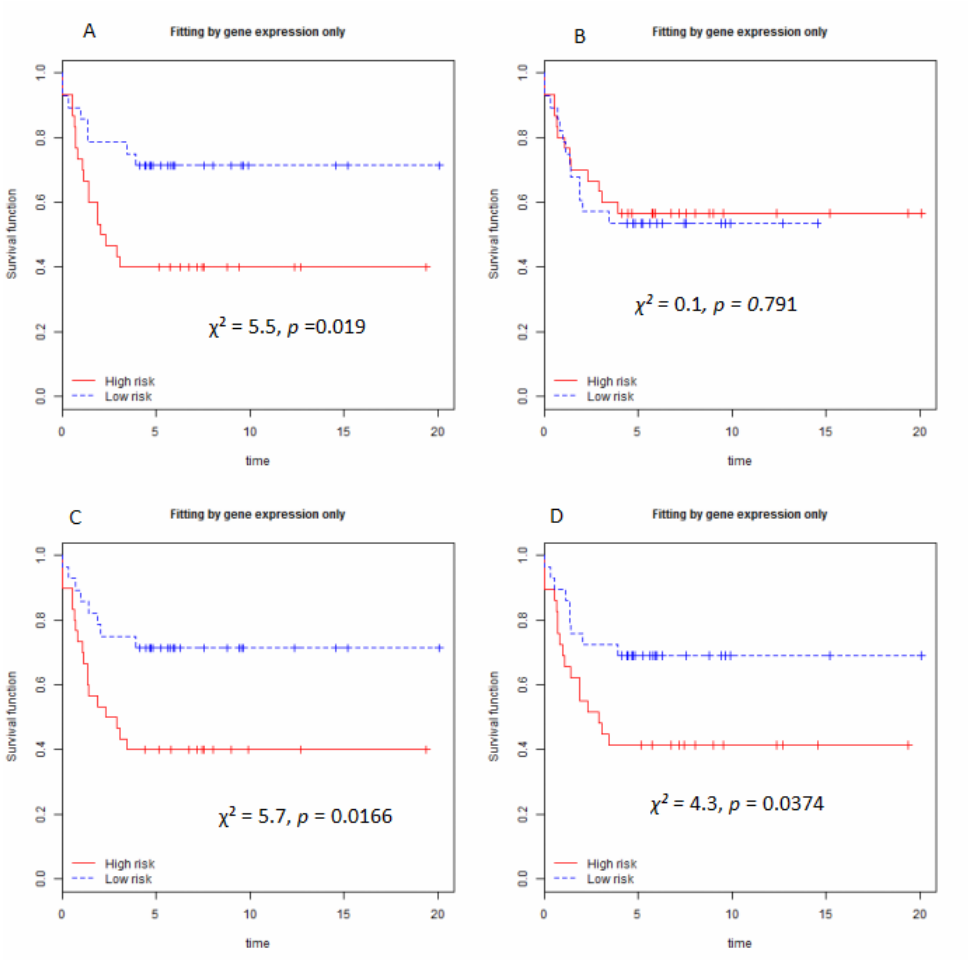

Figure 2.4: **Survival curves for the low and high risk groups** A: The survival curve generated by the SAM genelist, B: The survival curve generated by the LIMMA genelist, C: The survival curve generated by the LPC genelist and D: The survival curve generated by the QTA genelist.

of the genelists for a continuous outcome, and the prognostic properties of the genelists for DMFS. One simulated dataset and two publicly available datasets from melanoma studies are used in this regard. Results show that the selection of the DEGs heavily depends on the choice of the gene selection method. This may be due to the assumptions made by different methods. The LIMMA method assumes that the null distribution of the test statistics is the same for all genes. The QTA approach depends heavily on the normality and linearity assumptions, and the SAM method, in case of two groups scenario, assumes equal variance. Therefore, to obtain reliable results for detecting significant genes in microarray data analysis, we need to explore the characteristics of the data and then apply the most appropriate method under the given situation. Table 2.5 and 2.6 list the

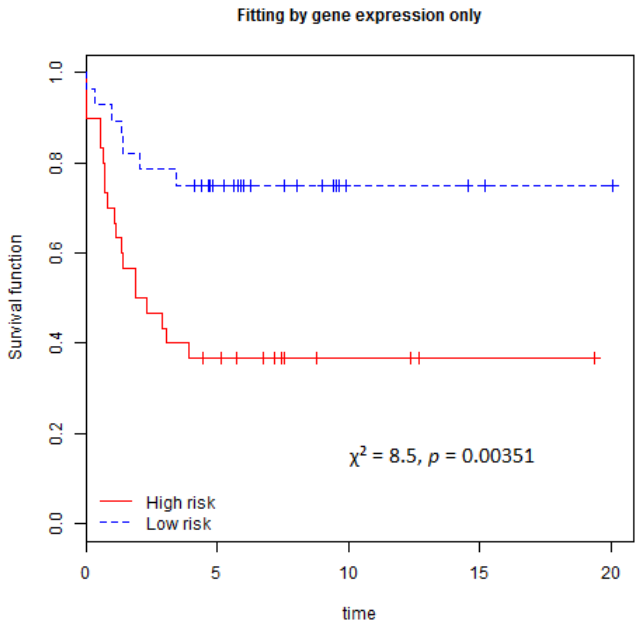

Figure 2.5: Survival curve for the low and high risk groups generated by the combined genelist containing 52 unique genes.

merits and the demerits of the four methods evaluated in this chapter.

In addition to finding DEGs, it is also important to assess the biological and clinical importance of these genelists. One way of doing this is by identifying gene signatures that are better predictors of a quantitative outcome or a patient's survival. This may help in tailoring therapeutic strategies to a single patient rather than the one-size-fits-all paradigm. Results from this chapter's work has shown that a combined genelists is more accurate in separating melanoma patients into high/low risk groups for developing distant metastasis. While the SAM approach was more powerful in terms of the number of significant genes detected using real dataset, the genelist generated by the QTA approach performed better in terms of prediction. Therefore, the QTA method would be preferred over the other approaches in predicting a quantitative outcome.

Omolo et al. (2013) employed the QTA method (together with a Bayesian procedure) to identify 165 genes that were associated with the G2 checkpoint function in melanoma lines. Some of these genes were found to be expressed differentially in wild-type (WT), NRAS-mutant and BRAF-mutant melanoma lines, through RNA expression analysis. This 165-list was also prognostic for distant metastasis-free survival in primary melanomas. Our SAM-list ($n = 33$), LIMMA-list ($n = 22$), LPC-list ($n = 7$) and QTA-

list ($n = 4$) had ten (10), three (3), three (3) and one (1) genes in common with the 165 gene list, respectively. Kaufmann et al. (2014) showed that some of the genes correlated with chromosomal instability ($n = 190$), obtained using the QTA and a Bayesian method, were linked to amplification or deletion of the gene, e.g. *DDR2*. Our SAM-list and the LIMMA-list had two (2) genes each in common with the 190-list, which included *DDR2*. Thus, some of the DEGs by the proposed statistical methods in this manuscript have been biologically validated to be true positives (TP) in recent studies.

Heterogeneity of the results in this chapter motivates the development of better methods that are more robust. In the next chapter, we will develop an alternative method, that is based on copula models, for finding differentially expressed genes when the outcome of interest is quantitative in nature.

Table 2.5: *Merits and demerits of the SAM, the LIMMA, the QTA and the LPC methods*

Method	Merits	Demerits
SAM	• Avoids small variance problem • Useful for small sample sizes • Avoid strong parametric assumption about the distribution of the test statistics • Does not involve any complex estimation procedure, (Only order statistics are involved)	• Not consistently performing well for small sample sizes • Sample variance correction technique is not model -motivated.
LIMMA	• Uses empirical Bayes model to correct sample variance • Can be adapted to many different and complex situations • Always yields good (BLUE) estimates • If assumptions are true, it provides a basis for inference	• Reliance on normal theory • Can't fit linear mixed models • Can't handle multiple levels of technical replication • If assumptions do not hold, conclusions are not to be trusted

Table 2.6: ***Cont.*** *Merits and demerits of the SAM, the LIMMA, the QTA and the LPC methods*

Method	Merits	Demerits
LPC	• Combine information across the genes • Can be applied to a variety of types of data • It improves many existing methods for the identification of significant features	• It does not rank genes using metric that is relevant or truly of interest • Assumes that gene work together in a pathway to cause an outcome
QTA	• Very simple to implement • Very powerful if the normality assumption is met • Flexible, can be extended to model more complex biological processes	• Valid only for linear relationship between variables • Less powerful for small sample sizes • Has a strong normality assumption

Chapter 3

Research Methodology

To appreciate the application of copulas to differential gene expression analysis, a basic understanding of copula theory is essential. This chapter introduces the concept of copulas, their properties and families. Methods of estimating copula parameters are also discussed in detail.

3.1 Introduction to Copulas

Theorem (Sklar 1959)**.** *Let F be an m-dimensional distribution function with margins $F_1, F_2, ..., F_m$. Then there exists an m-copula C such that for all x in $\bar{\mathbb{R}}^m$,*

$$F(x_1, x_2, ..., x_m) = C(F_1(x_1), F_2(x_2), ..., F_m(x_m)). \tag{3.1}$$

If $F_1, F_2, ..., F_m$ are all continuous, then C is unique; otherwise, C is uniquely determined on Ran $F_1 \times$ Ran $F_2 \times, ..., \times$Ran F_m . Conversely, if C is an m*-copula and $F_1, F_2, ..., F_m$ are distribution functions, then the function H defined by* (3.1) *is an m-dimensional distribution function with margins $F_1, F_2, ..., F_m$.*

Thus, a copula is a multivariate distribution on the m-dimensional unit cube, $[0, 1]^m$, with uniform marginals.

3.1.1 Probabilistic Interpretation of Copula Function

From the Sklar's theorem, copulas are joint distribution functions of standard uniform random variates:

$$C(u_1, ..., u_m) = Pr(U_1 \leq u_1, ..., U_m \leq u_m), \tag{3.2}$$

for any $\mathbf{u} = (u_1, ..., u_m)' \in [0, 1]^m$. We know that the probability integral transform of random variable $X_i \to F_i(x_i)$, is distributed as standard uniform $U_i, i = 1, ..., m$, that is,

$F_i(x_i) \sim U_i$. Then

$$\begin{aligned} C(F_1(x_1), ..., F_m(x_m)) &= Pr\{U_1 \leq F_1(x_1), ..., U_m \leq F_m(x_m)\} \\ &= Pr\{F_1^{-1}(U_1) \leq x_1, ..., F_m^{-1}(U_m) \leq x_m\} \\ &= Pr\{X_1 \leq x_1, ..., X_m \leq x_m\} \\ &= F(x_1, ..., x_m). \end{aligned} \tag{3.3}$$

3.2 Classes of Copulas

The commonly used classes of copula are the Archimedean and elliptical copulas (Yan, 2007). Figure 3.1 shows the scatter plots for 2000 samples generated from four different copulas namely: the Normal copula, the Frank copula, the Gumbel copula, and the Clayton copula. The standardized correlation matrix is used to determine the dependence structure of a copula since copulas are invariant to monotonic transformation of the margins. Some of the commonly used dispersion structures are: *exchangeable (ex)*, *Toeplitz (toep)*, *autoregressive of order 1 (ar1)*, and *unstructured (un)*. Correlation matrices corresponding to the mentioned structures are as follows for the case of $m = 3$ (Yan, 2007):

$$\begin{pmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 & \rho_1^2 \\ \rho_1 & 1 & \rho_1 \\ \rho_1^2 & \rho_1 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_3 \\ \rho_2 & \rho_3 & 1 \end{pmatrix} \tag{3.4}$$

respectively.

3.2.1 Independence Copula

The simplest copula function is the product copula that has the following form

$$C(u_1, u_2) = u_1 u_2, \tag{3.5}$$

where u_1 and u_2 are uniformly distributed over [0,1]. This copula corresponds to the independence case.

3.2.2 Archimedean Copulas

These types of copulas are common in applications because they are easy to construct and a great variety of copulas belong to this class. These copulas also posses nice properties

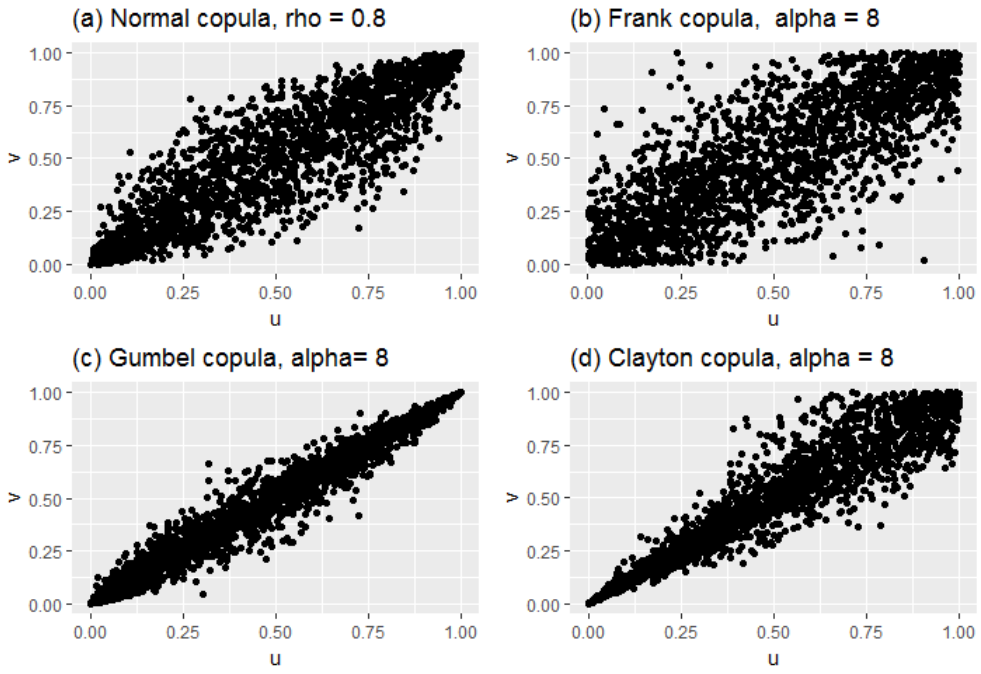

Figure 3.1: Scatter plots of 2000 random samples generated from the bivariate (a) Normal copula, (b) Frank copula, (c) Gumbel copula and (d) Clayton copula.

for example, most but not all extend to higher dimensions via the associativity property (Nelsen, 2006).

Definition: C is a bivariate Archimedean copula if it can be presented as

$$C(u, v) = \psi^{-1}[\psi(u) + \psi(v)], \tag{3.6}$$

where ψ is a continuous, strictly decreasing, convex function from $[0, 1]$ to $[0, \infty]$ such that $\psi(1) = 0$. The function ψ is called the generator function of the copula. The *pseudo-inverse* of ψ is the function $\psi^{[-1]}$ with $\text{Dom}\psi^{[-1]} = [0, \infty]$ and $\text{Ran}\psi^{[-1]} = [0, 1]$ given by (Nelsen, 2006)

$$\psi^{[-1]}(t) = \begin{cases} \psi^{-1}(t) & \text{if } 0 \leq t \leq \psi(0) \\ 0 & \text{if } \psi(0) \leq t \leq \infty \end{cases}$$

Table 3.1 highlights some of the commonly used bivariate Archimedean copulas with corresponding parameter ranges and generators. In the Clayton copula, the margins become independent as θ approaches zero. The downside of this copula is that it cannot account for a negative dependence. Likewise, the Frank copula attains independence as θ reaches zero. The Frank copula is symmetric in both tails. It is very popular because it can account for both negative and positive dependence. Just like the Clayton copula,

the Gumbel copula does not account for negative dependence.

Table 3.1: *Archimedean copulas and their generators*

Copulas	$c(u_1, u_2; \theta)$	Parameter θ range	Generator
Clayton	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{1/\theta}$	(0,∞)	$\frac{1}{\theta}(t^\theta - 1)$
Frank	$\frac{1}{\theta}\left(1 + \frac{(e^{\theta u_1}-1)(e^{\theta u_2}-1)}{e^\theta - 1}\right)$	$(-\infty, \infty)$	$-ln(\frac{e^{\theta t}-1}{e^\theta - 1})$
Gumbel	$\exp\{-[(-ln\ u_1)^\theta + (-ln\ u_2)^\theta]^{1/\theta}\}$	$[1, \infty)$	$(-ln\ t)^\theta$

There are many different copula functions belonging to the Archimedean family and a lot of different families or classes of copula functions but they are not commonly used in practical applications because of their analytical complexity. Nelsen (2006) gives an extensive review of these copulas.

3.2.3 Elliptical Copulas

An elliptical copula is the copula corresponding to an elliptical distribution. A general discussion about the elliptical distributions can be found in Fang et al. (1990). Let F be the multivariate cumulative distribution function (CDF) of an elliptical distribution. Let F_i be the CDF of the i^{th} marginal density and F_i^{-1} be its inverse function (quantile function), $i = 1, ..., m$. The elliptical copula determined by F is

$$C(u_1, ..., u_m) = F[F_1^{-1}(u_1), ..., F_m^{-1}(u_m)]. \tag{3.7}$$

Two copulas belong to this family: The Gaussian and the student t-copula
The Gaussian copulas: A bivariate normal copula is expressed as

$$C(u_1, u_2) = \Phi_\theta(\Phi^{-1}(u_1), \Phi^{-1}(u_2)), \tag{3.8}$$

where

$$\Phi_\theta = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left[-\frac{x^2 - 2\theta xy + y^2}{2(1-\theta^2)}\right] \mathrm{d}x\mathrm{d}y \tag{3.9}$$

is the standardized bivariate normal distribution function with correlation θ and

$$\Phi(u_1) = \int_{-\infty}^{\Phi^{-1}(u_1)} \frac{1}{\sqrt{2}\pi} \exp\left[-\frac{1}{2}x^2\right] \mathrm{d}x \tag{3.10}$$

denotes the univariate standardized distribution function. A normal copula allows for equal degrees of positive and negative dependence. This makes it flexible in applications. This study adopts this copula in most of its analysis.

The student-t Copulas: A bivariate student-t copula is expressed as

$$c^t(u_1, u_2) = \int_{-\infty}^{t_{\theta_1}^{-1}u_1} \int_{-\infty}^{t_{\theta_2}^{-1}u_2} \frac{1}{2\pi(1-\theta_2^2)^{1/2}} \times \left[1 + \frac{x^2 - 2\theta_2 xyt + y^2}{\nu(1-\theta_2^2)}\right]^{(-\theta_1+2)/2}, \quad (3.11)$$

where $t_{\theta_1}^{-1}(u_1)$ is the inverse of the CDF of standard univariate student-t distribution with θ_1 degree of freedom. θ_1 controls the heaviness of the tails. As $\theta_1 \to \infty$, the student-t copula behaves like the Gaussian copula.

3.3 Estimation of Copula Functions

This section describes different approaches for estimating copula functions. Given a copula density function, one can fit a copula model by estimating its parameters. A number of methods have been proposed in the literature. Most of the estimation methods proposed in the literature (Yan, 2007) are likelihood-based and include the exact maximum likelihood methods (EML), the inference for margins approach (IFM) and the canonical maximum likelihood estimation approach (CMLE). IFM and CMLE are also called multistage estimation methods. Like in any estimation problem, a model needs to be specified. The model of interest for all the above mentioned estimation approaches is expressed as

$$F(x_1, x_2, ..., x_m) = C(F_1(x_1), F_2(x_2), ..., F_m(x_m)), \quad (3.12)$$

where F is a continuous marginal with density f.

Model (3.12) may either be a parametric or semi-parametric model, depending on whether assumptions are made on the marginal distribution F or not. It is a fully parametric model if distribution assumption is made on the marginals. The estimation of parametric models relies on the assumption of parametric univariate maginal distribution. The success of estimating a parametric model depends on using an appropriate marginal distribution. Finding an appropriate marginal distribution is always not straight forward especially if the marginals show evidence of heavy tails and skewness. In semi-parametric models, no assumption is made on the marginals but the dependence structures, which in copulas, is assumed to come from some parametric family.

3.3.1 Copula density and likelihood function

From Sklar's theorem

$$F(x_1, x_2, ..., x_m) = C(F_1(x_1), F_2(x_2), ..., F_m(x_m)), \quad (3.13)$$

for $(x_1, ..., x_m)$ in support of F. Upon differentiation, (3.13) becomes

$$f(x_1, x_2, ..., x_m) = \frac{\partial^m C(F_1(x_1), ..., F_m(x_m))}{\partial F_1(x_1), ..., \partial F_m(x_m)} \prod_{i=1}^{m} \frac{\mathrm{d}F_i(x_i)}{\mathrm{d}x_i}$$
$$= c(F_1(x_1), ..., F_m(x_m)) \prod_{i=1}^{m} f_i(x_i) \tag{3.14}$$

Here, f, c and f_i are the densities for F, C and F_i, $i = 1, 2, ..., m$, respectively. Now, consider a random sample $\{(X_{1j}, ..., X_{mj});\ j = 1, ..., n\}$ from the distribution $F(x_1, ..., x_m)$,

$$L_n = \prod_{j=1}^{n} f(x_{1j}, ..., x_{mj}) = \prod_{j=1}^{n} c(F_1(x_{1j}), ..., F_m(x_{mj})) \prod_{j=1}^{n} \prod_{i=1}^{m} f_i(x_{ij}), \tag{3.15}$$

where L_n is the likelihood function. In practice, it is more convenient to work with the logarithm of the likelihood function because it simplifies subsequent mathematical analyses. Since the logarithm is monotonically increasing function, maximizing the log of a function is the same as maximizing the function itself. The log-likelihood representation of equation (3.15) is given as

$$\ell_n = \sum_{j=1}^{n} log\, c(F_1(x_{1j})..., F_m(x_{mj})) + \sum_{j=1}^{n} \sum_{i=1}^{m} log(f_i(x_{ij})). \tag{3.16}$$

3.3.2 Exact Maximum Likelihood Estimation Method

This approach estimates the copula parameter and the parameters of uncorrelated margins simultaneously. The estimation of marginals affects the estimation of the copula, and vice versa. The computation will also be of concern if both copula and marginals take some complicated form. The number of parameters to be estimated simultaneously can be large hence a computational burden.

Consider a copula-based parametric model for the random vector X, with cumulative distribution function

$$F(x_1, x_2, ..., x_m; \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, ..., \boldsymbol{\beta_m}) = C(F_1(x_1; \boldsymbol{\beta_1}), F_2(x_2; \boldsymbol{\beta_2}), ..., F_m(x_m; \boldsymbol{\beta_m}); \boldsymbol{\rho}), \tag{3.17}$$

where $F_i, ..., F_m$ are univariate cumulative distribution functions with respective parameters $\boldsymbol{\beta_1}, ..., \boldsymbol{\beta_m}$, and C is a family of copulas parametrized by a vector of parameters $\boldsymbol{\rho}$. For a sample of size n, the log likelihood function takes the form

$$\ell_n(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log c(F_1(x_{1j}; \boldsymbol{\beta_1}), ..., F_m(x_{mj}; \boldsymbol{\beta_m}); \boldsymbol{\rho}) + \sum_{j=1}^{n} \sum_{i=1}^{m} log(f_i(x_{ij}; \boldsymbol{\beta_i})), \tag{3.18}$$

The maximum likelihood estimator of $\boldsymbol{\theta}$ is given as

$$\hat{\boldsymbol{\theta}}_{EML} = \underset{\theta}{\operatorname{argmax}}\ \ell_n(\boldsymbol{\theta}), \tag{3.19}$$

where $\hat{\boldsymbol{\theta}}_{EML} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}})$, $\hat{\boldsymbol{\rho}}$ denotes a vector of estimates for the copula parameters, and $\hat{\boldsymbol{\beta}}$ denotes a vector of estimates for the parameters of the marginal distributions. We assume that the usual regularity conditions (Shao, 2003; Serfling, 2002) for asymptotic maximum likelihood theory hold for the multivariate model (that is the copula) as well as for all of its margins (the univariate probability density functions). Under the usual regularity conditions, the maximum likelihood estimator is asymptotically multivariate normal;

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{EML} - \theta_0) \to N(0, \mathfrak{F}^{-}(\theta_0)). \tag{3.20}$$

$\mathfrak{F}^{-}(\theta_0)$ is the Fisher's information matrix and θ_0 is the true value. The covariance matrix of $\hat{\boldsymbol{\theta}}_{EML}$ (Fisher's information matrix) may be estimated by the inverse of the negative Hessian matrix of the likelihood function.

3.3.3 Inference Function for Margins Method

This method was proposed in a general framework in Xu (1996) and is discussed for copula in Joe (2005). It is motivated by the fact that EML estimation approach is computationally intensive as the number of parameters to be estimated increases.

In this approach, the log-likelihood function is maximized in two stages. In the first stage, the $i = 1, ..., m$ log-likelihood functions of the margins are optimized to obtain the estimates for the parameters of the margins $\boldsymbol{\beta_i}$, $i = 1, ..., m$.

$$\hat{\boldsymbol{\beta}}_i = \sum_{j=1}^{n} log(f_i(x_{ij}; \boldsymbol{\beta_i})). \tag{3.21}$$

The vector of the copula parameters are then estimated in the second stage using the previous estimators, $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_m)$, as follows:

$$\hat{\boldsymbol{\theta}}_{IFM} = \underset{\theta}{\operatorname{argmax}} \sum_{j=1}^{n} \log c(F_1(x_{1j}; \hat{\boldsymbol{\beta}_1}), ..., F_m(x_{mj}; \hat{\boldsymbol{\beta}_m}); \boldsymbol{\theta}). \tag{3.22}$$

The efficiency of this approach was studied in Joe (2005). Under regular conditions, $\hat{\boldsymbol{\theta}}_{IFM}$ is asymptotically normal;

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{IFM} - \theta_0) \to N(0, \mathfrak{G}^{-}(\theta_0)), \tag{3.23}$$

where $\mathfrak{G}^{-}(\theta_0)$ is the Godambe information matrix. This approach yields parameters that are less efficient in the presence of strong association than the EML approach though the loss of efficiency is not great (Joe, 2005).

3.3.4 Canonical Maximum Likelihood Estimation(CMLE) Method

In this approach, no parametric assumptions are made on the marginals and therefore, it relies on the concept of empirical marginal transformation. The transformation approximates the unknown parametric marginal $F_i(x_i)$ with an empirical distribution function $\hat{F}_i(x_i)$ given by

$$\hat{F}_i(x_i) = \frac{n}{n+1}\frac{1}{n}\sum_{j=1}^{n} I\left(X_{ij} \leq x_i\right), \tag{3.24}$$

where I is the indicator function. Rescaling the empirical distribution by $\frac{n}{n+1}$ avoids the potential unboundedness of $\log(c(F_1(x_{1j}), ..., F_m(x_{mj}); \boldsymbol{\theta})$ as some of the $F_i(x_{ij})$'s tend to be one (Genest et al., 1995). These empirical CDFs are then used in (3.15) to estimate the copula parameter by optimizing the pseudo-loglikelihood function as follows:

$$\hat{\boldsymbol{\theta}}_{CML} = \underset{\theta}{\operatorname{argmax}} \sum_{j=1}^{n} \log c(\hat{F}_1(x_{1j}), ..., \hat{F}_m(x_{mj}); \boldsymbol{\theta}) \tag{3.25}$$

Under suitable regularity conditions, $\hat{\boldsymbol{\theta}}_{CML}$ is consistent and is asymptotically normal Genest et al. (1995).

3.4 Copula-based Dependence Measure

In this section, the two most widely known scale-invariant measures of association, both of which measure a form of concordance, are briefly described. These are the population Kendall's tau and the Spearman's rho. Other existing measures of concordance are Gini's gamma and Blomqvist's beta.

3.4.1 Concordance

Let (x_i, y_i) and (x_j, y_j) denote two observations from a vector (X, Y) of continuous random variables. We say that (x_i, y_i) and (x_j, y_j) are concordant if $x_i < x_j$ and $y_i < y_j$, or if $x_i > x_j$ and $y_i > y_j$. Similarly, we say that (x_i, y_i) and (x_j, y_j) are discordant if $x_i < x_j$ and $y_i > y_j$ or if $x_i > x_j$ and $y_i < y_j$. Alternatively, (x_i, y_i) and (x_j, y_j) are concordant if $(x_i - x_j)(y_i - y_j) > 0$ and discordant if $(x_i - x_j)(y_i - y_j) < 0$ (Nelsen, 2006).

3.4.2 Kendall's Tau

The definition of Kendall's tau as highlighted in Nelsen (2006) is as follows.

Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed random vectors each with joint distribution function H. Kendall's tau is defined as the difference between the probabilities of concordance and discordance:

$$\tau_{X,Y} = Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - Pr[(X_1 - X_2)(Y_1 - Y_2) < 0] \tag{3.26}$$

This can be expressed in terms of Copula as follows

$$\tau_{X,Y} = \tau_C = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1 \tag{3.27}$$

where C is the copula associated to (X, Y).

3.4.3 Spearman's Rho

This measure is also based on concordance and discordance. To obtain this measure, we let (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) be three independent random vectors with a common joint distribution function H whose margins are F and G and copula C. Spearman's rho is the probability of concordance minus the probability of discordance for two vectors (X_1, Y_1), (X_2, Y_3) (Nelsen, 2006) and it is expressed as

$$\rho_{X,Y} = 3(Pr[(X_1 - X_2)(Y_1 - Y_3) > 0] - Pr[(X_1 - X_2)(Y_1 - Y_3) < 0]) \tag{3.28}$$

The pair (X_3, Y_2) could be used as well. We note that while the joint distribution function of (X_1, Y_1) is $H(x, y)$, the joint distribution function of (X_2, Y_3) is $F(x)G(y)$ because X_2 and Y_3 are independent.

In terms of copulas, (3.28) can be expressed as

$$\rho_{X,Y} = \rho_C = 12 \int_0^1 \int_0^1 (C(u, v) du dv - 3. \tag{3.29}$$

Equation (3.29) can also be expressed as

$$\rho_{X,Y} = \rho_C = 12 \int_0^1 \int_0^1 [C(u, v) - uv] du dv. \tag{3.30}$$

3.5 Choosing A Copula

There exists a wide range of copula families that are available for use. For detailed coverage about copula families, see Nelsen (2006). This presents a challenge when it comes to specification of a suitable copula for a given dataset.

Several tests have been proposed for the copula specification. The most commonly used is the goodness-of-fit test (Fermanian, 2005; Wang, 2010; Genest et al., 2006; Dobri and Schmid, 2007; Berg, 2009). Goodness-of-fit tests are based on a direct comparison of the dependence implied by the copula with the dependence observed in the data.

In most empirical applications, the unique copula C is assumed to come from a parametric family $C_0 = \{C_\theta, \theta \in \Theta\}$ with $\Theta \subset R$. In goodness-of-fit testing for copula models, the hypothesis of interest is given by $H_0 : C \subset C_0$, i.e. that the copula C belongs to a predetermined parametric family C_0. For testing H_0, the marginal distributions are treated as nuisance parameters and are replaced by their empirical distribution functions $\hat{F}_i(x_i)$ as defined in (3.24) Genest et al. (2009).

Copulas can also be selected according to their ranks based on some criteria. The most commonly used are the Akaike's Information Criteria (AIC) Akaike (1974) and the Bayesian's Information Criteria (BIC) Schwarz (1978). These are defined as follows:

$$AIC = -2\sum_{j=1}^{n} \ln[c(u_{1j}, u_{2j}); \theta] + 2K. \tag{3.31}$$

$$BIC = -2\sum_{j=1}^{n} \ln[c(u_{1j}, u_{2j}); \theta] + K\ln(n). \tag{3.32}$$

Here, $u_{ij} = F_i(x_{ij}),\ i = 1, 2$, and $K = 1$ for the one-parametric copulas. Similarly, $K = 2$ for the two-parametric copulas. A copula with the least AIC or BIC is chosen to be the best.

3.6 Application of Copula in Different Fields

Copula models have become popular modelling tools in many fields where the main interest is in the dependence of marginal distribution. Copula methods have mainly been applied in finance and actuarial science (Romano, 2002; Cherubini et al., 2004). In bioinformatics, Owzar et al. (2007) used copulas to identify prognostic genes. Kim et al. (2008) reconstructed gene networks from gene expression data using copulas. Bao et al. (2009) described a semiparametric copula model via extended rank likelihood which allows estimation of the dependence structure of multiple continuous variables, ordinal

variables, or the mixture of those two types. Yuan et al. (2008) proposed a semiparametric copula method for microarray-SNP genomewide association analysis using pedigree data. They performed the gene copy family analysis using a multivariate normal copula. Li et al. (2006) developed and implemented a copula variance-components (VC) method, that directly models the nonnormal distribution using Gaussian copulas. Escarela and Carrire (2003) proposed a fully parametric model for the analysis of competing risks data where the types of failure may not be independent. They applied their copula method to a prostate cancer data set. Copulas has also been applied in the energy sector. Louie (2014) modeled wind power using Archimedean and Gaussian copula.

Copula methods have been widely applied in bioinformatics but their applications in microarray data for gene selection is still limited (Owzar et al., 2007; Bao et al., 2009; Emura and Chen, 2016). In the next chapter, we use the concept of copula models discussed in this chapter to find DEGs when we have a quantitative outcome.

Chapter 4

Using Copulas to Select Differentially Expressed Genes

In this chapter, we develop a copula-based algorithm for finding differentially expressed genes. The copula-based algorithm so developed is evaluated using several simulation datasets. It is then applied to a real dataset to identify prognostic genes. This chapter is organised as follows: In section 4.1, we provide the motivation for the development of the copula-based algorithm. In section 4.2, we develop an algorithm based on the copula models for differential gene expression. Section 4.3, describes the simulation set up for evaluating the copula-based approach and the application is described in section 4.4.

4.1 Motivation

Melanoma of the skin is among the most common cancer types in the United States. It is the fifth and seventh most commonly diagnosed carcinoma in men and women, respectively (Siegel et al., 2017). In Kenya, data on melanoma of the skin is not well documented. According to the information from HealthGrove (2017), the annual mortality rate per 100,000 people from malignant skin melanoma in Kenya is 20%. A major challenge with melanoma is the identification of therapeutic targets. Multi-gene signatures have shown promise in this regard and a number of these signatures have been developed within the last decade (Winnepenninckx et al., 2006; Mandruzzato et al., 2006; John et al., 2008; Bogunovic et al., 2009; Jönsson et al., 2010; Carson et al., 2012; Omolo et al., 2013; Kaufmann et al., 2014). Winnepenninckx et al. (2006) identified 254 genes that were associated with distant metastasis-free survival of patients with primary melanoma, of which 174 correspond to known genes. Mandruzzato et al. (2006) identified 80 probes that were correlated with overall survival in a cohort of patients with stage III and IV melanoma, 30 of which were associated with survival. John et al. (2008) found 21 differ-

entially expressed genes which showed ability to separate prognostic groups. Bogunovic et al. (2009) identified a group of 266 genes which can predict survival in metastatic melanoma. Jönsson et al. (2010) identified gene signatures that were associated with four distinct subtypes of metastatic melanoma (immune response, pigmentation differentiation, proliferation, and stromal composition genes). They found few common genes between the Winnepenninckx et al. (2006) metastasis signature and their proliferative subtype. Carson et al. (2012) identified 316 probes whose expression was correlated with G_1 checkpoint function in melanoma lines. When applied to microarray data from primary melanomas, the 316 probe list was prognostic of 4-year distant metastasis-free survival. Omolo et al. (2013) identified 165 genes that were correlated with G_2 checkpoint function, 32 of which were prognostic. The signature was enriched in lysosomal genes and contained numerous genes that are associated with regulation of chromatin structure and cell cycle progression. Kaufmann et al. (2014) generated a gene signature with 190 genes which were correlated with chromosomal instability index (CIN). This gene signature was however found not to be prognostic of metastasis-free survival.

The development of such gene signatures require use of statistical methods. Carson et al. (2012); Omolo et al. (2013) and Kaufmann et al. (2014) used parametric methods based on the t-test with multiple corrections. One advantage of these methods is they offer a straightforward approach to calculating p-values and confidence intervals. Moreover, for large samples, the distribution of the t-statistic is independent of the overall expression level of the gene. Unfortunately, for small sample sizes, the t-test based methods depend on strong parametric assumptions. These assumptions may be violated in practice, and so non-parametric methods have also been applied in some studies (Mandruzzato et al., 2006; Bogunovic et al., 2009; John et al., 2008). For these methods, the distribution of random errors are estimated without strong parametric assumptions. The Significance analysis of microarray (SAM) method (Tusher et al., 2001), in particular, avoids high variance that results from estimating the variance of each gene separately. When sample size is small, any method that reduces the variance in the estimates produce more accurate results. The non-parametric methods also have disadvantages which vary from one method to the other. For example, the Wilcoxson-test approach exhibits low power in the identification of differentially expressed genes (Troyanskaya et al., 2002). For a detailed review of methods for finding differentially expressed genes, see Troyanskaya et al. (2002); Bair (2013); Bandyopadhyay et al. (2014); Chaba et al. (2017).

Despite all the proposed methods mentioned above, there is no unanimous agreement on any particular gene selection method. Furthermore, most methods were developed for finding differentially expressed genes based on groups or classes of samples (discrete

covaiates). However, there are many outcomes of interest that are continuous in nature.

In this chapter, we propose an algorithm for selecting genes associated with a continuous but non-clinical outcome based on a semi-parametric copula model. An advantage of the copula-based approach is its compatibility with any distribution function. This allows for the relaxation of the assumption of specific distribution. Owzar et al. (2007) has applied a copula-based approach to identify genes that are differentially expressed between stage I and III lung cancer patients based on survival copulas and family wise error rate (FWER) control. In contrast, our proposed algorithm is based on a continuous outcome from melanoma cell lines and controls for the false discovery rate (FDR), since the FWER is often too conservative (Benjamini and Hochberg, 1995).

The performance of our copula-based approach in terms of power is assessed via simulations. The method is then applied to a melanoma cell lines dataset to find genes that are correlated with G_2 checkpoint function. The gene signature generated by copula approach is then subjected to an independent primary melanoma dataset to determine if it is prognostic of 4-year distant metastasis-free survival in melanoma patients.

4.2 Copula Model for Differential Gene Expression

We are interested in the pairwise correlation between each gene's expression profile and a quantitative outcome. Therefore, the copula of interest is the bivariate copula ($m = 2$). Suppose a microarray experiment consists of n subjects/samples and G genes. Let $\mathbf{x_i} = (x_{1i}, ..., x_{ni})'$ be a vector of gene expression profile for gene i and $\mathbf{y} = (y_1, ..., y_n)'$ be a vector of the covariate of interest (quantitative trait). We wish to find K genes that are correlated with Y, $0 < K < G$. That is, we are interested in determining whether, for each gene i, $\mathbf{x_i}$ and $\mathbf{y}$ are independent or not. The test for independence, thus, becomes testing for null hypothesis

$$H_{0i} : Y \perp X_i \ \ (X_i \text{ and Y are independent}) \tag{4.1}$$

against the alternative hypothesis

$$H_{1i} : Y \not\perp X_i \ \ (X_i \text{ and Y are independent}) \tag{4.2}$$

The biological questions of differential gene expression in microarray consists of multiple hypothesis testing problem in which several hypotheses are tested simultaneously. In this

case, the hypothesis of interest becomes

$$H_0 : Y \perp X_i \text{ for all i } = \bigcap_{i=1}^{G} H_{0i} \tag{4.3}$$

vs.

$$H_1 : Y \not\perp X_i \text{ for some i } = \bigcup_{i=1}^{G} H_{1i}. \tag{4.4}$$

In terms of copulas, assume that for each gene i, the joint distribution of Y and X_i is generated by a parametric copula $C(u_1, u_2; \theta_i)$ such that

$$H_i(y, x) = C[F(y), F_i(x); \theta_i], \tag{4.5}$$

where $H_i(y, x)$, $F(y)$ and $F_i(x)$ are the CDFs of (Y, X_i), Y and X_i respectively. Here $u_1 = F(y)$, $u_2 = F_i(x)$ and θ_i is the dependence parameter. Equation (4.3) and (4.4) now becomes

$$H_0 : \bigcap_{i=1}^{G} \left[C(u_1, u_2; \theta_i) = u_1 u_2 \text{ for all } (u_1, u_2)^T \in [0, 1]^2 \right], \tag{4.6}$$

vs.

$$H_1 : \bigcup_{i=1}^{G} \left[C(u_1, u_2; \theta_i) \neq u_1 u_2 \text{ for some } (u_1, u_2)^T \in [0, 1]^2 \right]. \tag{4.7}$$

A normal copula, for instance, attains independence when $\theta_i = 0$. In this case, the global hypothesis to test for the dependence in terms of θ_i is expressed as

$$H_0 : \bigcap_{i=1}^{G} (\theta_i = 0), \text{vs. } H_1 : \bigcup_{i=1}^{G} (\theta_i \neq 0). \tag{4.8}$$

4.2.1 Hypothesis Testing

We are testing (4.8), so G hypothesis tests are performed simultaneously. Each hypothesis tests $H_0 : \theta = 0$. We need to estimate the distribution of $\hat{\theta}_i$ under the null hypothesis. Rather than assuming a parametric distribution for the null hypothesis, a permutation resampling based approach (Westfall and Young, 1993) can be used to find a gene-specific p-value. For a given α, a gene is differentially expressed if its p-value $< \alpha$. Since the goal is to test several hypothesis simultaneously, it is crucial to employ a method that accounts for multiplicity. The false discovery rate (FDR) (Storey and Tibshirani, 2003) is used in this regard. The global null hypothesis (4.8) is rejected if at least one of its components (H_{0i}) is rejected, based on the estimated FDR values.

4.2.2 Copula Algorithm for Identifying DEGs

Our copula-based algorithm for finding differentially expressed genes (DEGs) can be summarised as follows:

1. Estimate θ_i using the CMLE method. In the CMLE approach, no assumption is made on the marginal distribution. The marginal distribution for each gene, $F_i(x_i)$ and a quantitative outcome, $F(y)$, are replaced with their estimators $\hat{F}_i(x_i)$ and $(\hat{F}(y))$, respectively, to obtain $\hat{\theta}_i$.

$$\hat{\theta}_{\mathbf{i}} \approx \arg\max \sum_{j=1}^{n} \log c(\hat{F}_i(x_i), \hat{F}(y); \theta). \tag{4.9}$$

   A detailed explanation of the CMLE method is in Section 3.3.4.

2. Find gene-specific p-values (unadjusted p-values) using the permutation based re-sampling method. Permutation approach provides an efficient method to testing when data do not conform to the distribution assumptions. To compute unadjusted p-value for each gene, we follow the procedure below.

   (a) Permutate the quantitative outcome column B times as you hold the gene expressions matrix fixed.

   (b) For the b^{th} permutation, $b = 1...B$, compute test statistics $\hat{\theta}_{1b}, ..., \hat{\theta}_{Gb}$ for each hypothesis using equation (4.9).

   (c) After the B permutations are done, for two-sided alternative hypotheses, the permutation p-value for hypothesis H_i is

$$p_i = \frac{\#\{b : |\hat{\theta}_{ib}| \geq |\hat{\theta}_i|\}}{B} \tag{4.10}$$

   where $\hat{\theta}_i$ is the original $\hat{\theta}$ for the i^{th} gene before the permutation.

3. Apply the FDR approach to control for type I error. The procedure below outline the steps followed in estimation of FDR given the p-value (Storey and Tibshirani, 2003).

   (a) Let $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(G)}$ be the ordered p-values. This also denotes the ordering of the features in terms of their evidence against the null hypothesis.

   (b) For a range of λ , say $\lambda = 0, 0.01, 0.02, ...$, 0.95, calculate

$$\hat{\pi}_0(\lambda) = \frac{\#(p_j > \lambda)}{G(1-\lambda)}.$$

(c) Let $\hat{f}$ be the natural cubic spline with 3 df of $\hat{\pi}_0(\lambda)$ on λ.

(d) Set the estimate of π_0 to be $\hat{\pi}_0 = \hat{f}(1)$.

(e) Calculate $\hat{q}(p_{(G)}) = \hat{\pi}_0 p_{(G)}$.

(f) For $i = G-1, G-2, ..., 1$, calculate

$$\hat{q}(p_{(i)}) = \min\left\{\frac{\hat{\pi}_0 \mathrm{G} p_{(i)}}{i}, \hat{q}(p_{(i+1)})\right\}.$$

(g) The estimated q-value for the i^{th} most significant feature is $\hat{q}(p_{(i)})$.

4. A gene is differentially expressed if its estimated FDR (estimated q-value) is less than some specified value $\alpha \in$ [0,1]

An R code for implementing the above algorithm is available in the Appendix C.

4.3 Simulated Gene Expression Data

Twelve simulation scenarios are considered in evaluating the performance of the proposed copula method in terms of power. Let n and G denote the number of samples and genes, respectively. Further, let D denote the number of genes assumed to be truly differentially expressed. Then $(G-D)$ genes are assumed to be non-differentially expressed. The gene expression data matrix, $\mathbf{X}$, is a $G \times n$ matrix of log2-ratios. We can write $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$, where $\mathbf{X_1}$ and $\mathbf{X_2}$ are $D \times n$ and $(G-D) \times n$ matrices, respectively. We set $D \in (50, 100, 200)$, $n \in (20, 35, 50, 100)$ and G to be 1000. We generate the $(1000 - D)$ genes from the standard normal distribution. To generate the D genes, we use the standard normal distribution in conjunction with the Cholesky decomposition (Golub and Van Loan, 1996) of their correlation matrix as follows:

1. Generate an unstructured correlation matrix $\mathbf{\Omega}$. $\mathbf{\Omega}$ is a $(D+1) \times (D+1)$ matrix that has $(i,j)^{th}$ element given by $\omega_{i,j} = \mathrm{corr}(x_i, x_j)$

2. Find the Cholesky factor, $\mathbf{A}$, of $\mathbf{\Omega}$ such that $\mathbf{\Omega} = \mathbf{A}\mathbf{A}'$.

3. Let $\mathbf{z}_i \sim N(0, \mathrm{I}_n), i = 1, 2, ..., (D+1)$.

4. $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{D+1})'$

5. $\mathbf{X}_{D+1} = \mathbf{A}\mathbf{Z}$

$\mathbf{X}_{D+1}$ is the gene expression matrix for D genes assumed to be differentially expressed and a covariate $\mathbf{y}$. $\mathbf{y}$ can take any of the $D+1$ row vectors from the matrix $\mathbf{X}_{D+1}$. $\mathbf{X}_1$ is therefore a submatrix of $\mathbf{X}_{D+1}$ with dimensions $D \times n$. This simulation set-up assumes that each gene is observed across each sample. The R code for the above simulation is available in the Appendix B.

The developed copula method is applied to the 12 simulated datasets to evaluate its power in identifying DEGs. We transpose $\mathbf{X}$ in the analysis, so that $\mathbf{X}'$ has genes on the columns and samples on the rows. We follow the procedure in section 4.2.2 to identify DEGs at different estimated FDR values. A normal copula is assumed. See Appendix A for description a normal copula. Power is calculated as the ratio of the number of correctly identified differentially expressed genes, true positives (TP), to the total number of actual DEGs, D. Thus,

$$\text{Power} = \frac{TP}{D} \tag{4.11}$$

4.4 Application

We apply the developed copula-based algorithm to a publicly available melanoma data. Gene expression profiles from the melanoma cell lines dataset with 3,860 genes and 35 samples is used. The G_2 checkpoint function data is used as the quantitative trait data. All these datasets are described in Chapter 2, Sub-Section 2.4.1.

A normal copula is assumed for the analysis of the melanoma dataset. Such an assumption was made in the Owzar et al. (2007) for lung cancer. Genes that are correlated with the G_2 checkpoint function are selected based on estimated FDR values. The predictive quality of the copula genelist is assessed via the least absolute shrinkage and selection operator (LASSO) algorithm (Tibshirani, 1996) with leave-one-out cross-validation (LOOCV) method. To check the biological significance of the G_2 signature generated by the copula method, we use the independent dataset in Winnepenninckx et al. (2006) (dataset described in Chapter 2, Section 2.4.1) to identify genes that could predict a patient's risk (low/high) for developing distant metastasis within 4 years of primary diagnosis. The supervised principal component method (Bair and Tibshirani, 2004) is used to separate the samples into high/low risk group. This procedure was implemented in the BRB-ArrayTools software.

4.5 Results and Discussion

Heatmaps of the simulated datasets with the 35 samples can be seen in Figure 4.1. From Table 4.1, we see that as the estimated FDR values increase, more genes are identified as being differentially expressed. For example, for $n = 35$, $D = 50$ and FDR = 0.05, 49 genes are identified and at FDR = 0.1 for the same n and D, 52 genes are identified. The same pattern is seen for the other values of n. Table 4.2 shows the power of the copula method at different estimated FDR levels for four sample sizes, $n = 20$, 35, 50 and 100. The results show that the power of the copula method is sensitive to low sample sizes. For example, the power is 0.58 when $n = 20$ at $D = 50$ but increases to 1 for the same value of D as n increases to 100. For the sample size of at least 35, the least value of the power observed from the analysis is 0.98. This shows that the copula method is quite powerful in finding differentially expressed genes. The copula approach is also robust to different sample sizes especially as the number of known DEGs increases.

Table 4.1: *DEGs at FDR level between 0.001 to 1 on twelve simulated dataets each with 1000 genes. Sample size was set at n = (20, 35, 50 and 100). Number of significant genes were set to be 50, 100 and 200.*

		Estimated FDR Threshold					
n	**D**	**0.001**	**0.01**	**0.025**	**0.05**	**0.1**	**0.2**
20	50	30	30	30	44	49	59
	100	79	92	95	103	117	132
	200	137	184	197	209	222	251
35	50	49	49	52	53	54	62
	100	99	101	101	105	110	121
	200	201	201	204	209	222	258
50	50	50	50	51	51	53	61
	100	100	101	102	106	112	124
	200	202	205	210	216	223	249
100	50	50	50	50	50	55	58
	100	100	101	101	107	111	136
	200	201	201	208	212	226	259

When applied to the cell lines dataset, the copula method identified 9 genes at FDR < 0.01 and 25 genes at FDR < 0.2 for G_2. Table 4.3 lists the genes that are correlated with G_2 checkpoint function at FDR < 0.2. We compare our results and the results

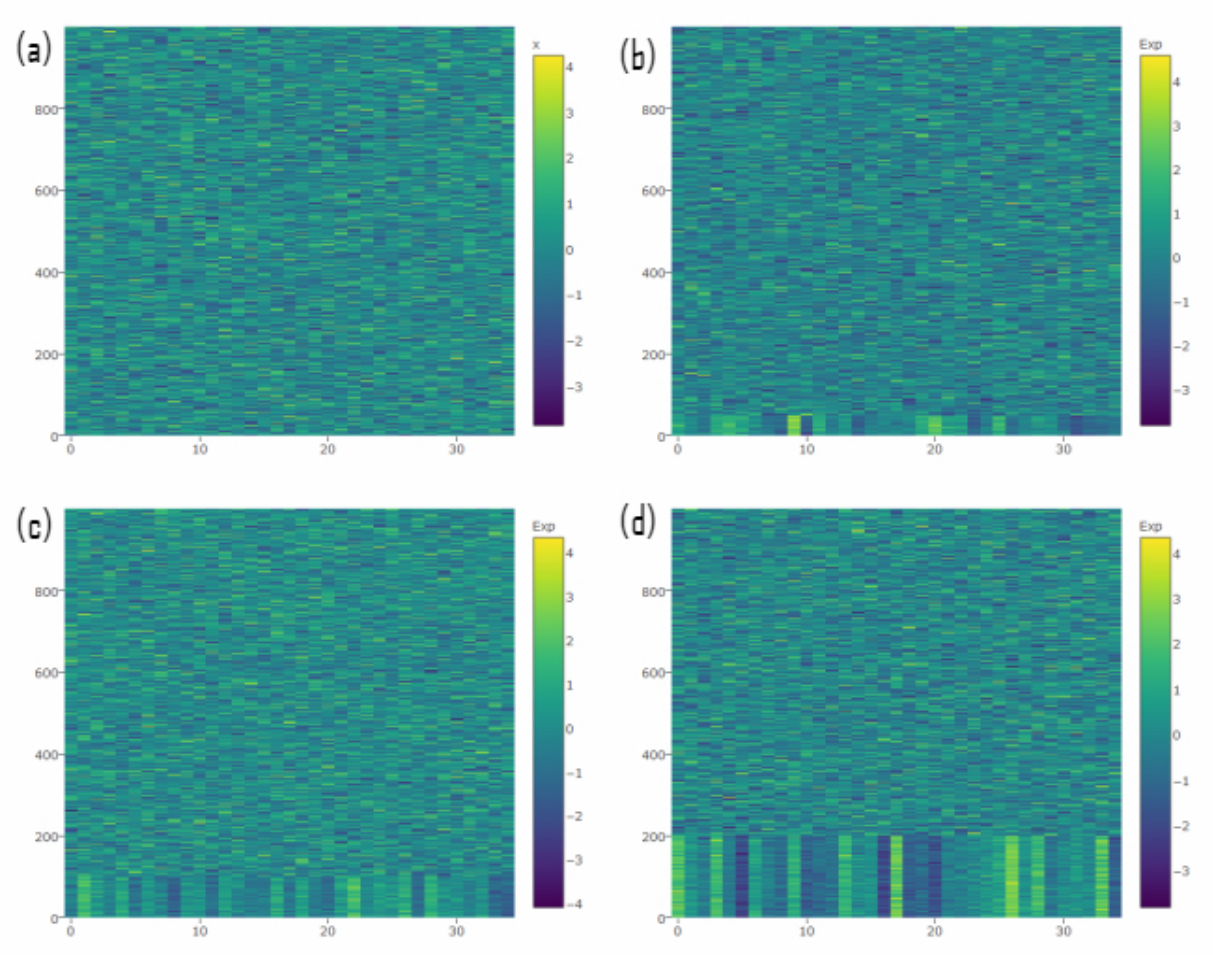

Figure 4.1: **Heatmaps of simulated data**. Simulated data contain 1000 genes and 35 samples. In (a) No assumption is made of the number of DEGs, in (b) 50 genes, in (c) 100 genes and in (d) 200 genes , are assumed to be correlated with quantitative outcome.

presented in the paper by Omolo et al. (2013). This paper also selected genes based on G_2 checkpoint function. They found 165 genes that were correlated with G_2. These 165 were unique genes generated by two methods; Bayesian and quantitative trait analysis (QTA). The intersection between our 25 genelist and their 165 genelist generated 3 genes namely *ZNF711, DGKE* and *ARNTL2*. The intersection results are in Figure 4.2. It is important to note that the QTA method applied in the paper by Omolo et al. (2013) did not adjust for multiplicity. Therefore, a direct comparison of the two genlists, 165 genelist and our 25 genelist, may not be appropriate.

The results from the LASSO analysis show that the copula genelist can be used to predict the G_2 checkpoint function ($r = 0.558$, p $= 0.004939$). The coefficient of determination was however low ($r^2 = 0.31$). We further subjected our genelist to a

Table 4.2: *Power at FDR level between 0.001 to 1 on six simulated datasets each with 1000 genes. Sample size was set at n = (20, 35, 50 and 100). Number of significant genes were set to be 50, 100 and 200.*

		Estimated FDR Threshold					
n	**D**	**0.001**	**0.01**	**0.025**	**0.05**	**0.1**	**0.2**
20	50	0.58	0.58	0.58	0.80	0.88	0.96
	100	0.79	0.92	0.95	0.98	0.99	1.00
	200	0.69	0.92	0.98	1.00	1.00	1.00
35	50	0.98	0.98	1.00	1.00	1.00	1.00
	100	0.99	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00
50	50	1.00	1.00	1.00	1.00	1.00	1.00
	100	1.00	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00
100	50	1.00	1.00	1.00	1.00	1.00	1.00
	100	1.00	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00

survival risk prediction analysis to asses its biological importance. Our list generated 4 prognostic genes which shows a significant separation of the samples into low and high risk group ($\chi^2 = 5.9$ p $= 0.0147$). See results in Figure 4.4. An almost similar results were seen in Omolo et al. (2013) for their 32 prognostic genes ($\chi^2 = 5.6$ p $= 0.018$). Our list of 25 genes performs better in SRP than any randomly selected 25 genes from the 3860 genes ($\chi^2 = 0.1$ p $= 0.655$). Unsupervised hierarchical clustering indicate no significant (p-value $= 0.317$) separation of the incidences of distance metastasis (**Fig.4.4**). Intersecting the prognostic genes from these two studies, only one gene, *ZNF711*, is generated. This gene, however, has not been previously reported in relation to melanoma development. It lies in a region of the X chromosome which has been associated with mental retardation (Tarpey et al., 2009).

4.6 Conclusion

In this chapter, we have proposed a copula-based algorithm for finding differentially expressed genes when the outcome of interest is continuous. In the proposal, a normal copula is employed in the analysis. We have shown the potential of the proposed

Table 4.3: *List of genes that are correlated with G_2 checkpoint function as selected by the copula approach at FDR < 0.2*

Agilent ID	Symbol	Gene Name
A_23_P14612	FGF7	fibroblast growth factor 7(FGF7)
A_23_P153964	INHBB	inhibin beta B subunit(INHBB)
A_23_P203115	TMEM25	transmembrane protein 25(TMEM25)
A_23_P211631	FBLN1	fibulin 1(FBLN1)
A_23_P214080	EGR1	early growth response 1(EGR1)
A_23_P217297	ZNF711	zinc finger protein 711(ZNF711)
A_23_P364504	ERFE	erythroferrone(ERFE)
A_23_P369328	C10orf35	chromosome 10 open reading frame 35(C10orf35)
A_23_P389250	Smco2	single-pass membrane protein with coiled-coil domains 2(SMCO2)
A_23_P393034	HAS3	hyaluronan synthase 3(HAS3)
A_23_P69537	NMU	neuromedin U(NMU)
A_24_P130952	MLK4	mixed lineage kinase 4(MLK4)
A_24_P196665	GNGT1	G protein subunit gamma transducin 1(GNGT1)
A_24_P20814	KHDC1L	KH domain containing 1 like(KHDC1L)
A_32_P209230	CITED4	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 4(CITED4)
A_32_P232559	PRKCQ-AS1	PRKCQ antisense RNA 1(PRKCQ-AS1)
A_32_P399546	ARNTL2	aryl hydrocarbon receptor nuclear translocator like 2(ARNTL2)
A_32_P540991	DGKE	diacylglycerol kinase epsilon(DGKE)
A_23_P153958	Unknown	Unknown
A_32_P134427	Unknown	Unknown
A_32_P154726	Unknown	Unknown
A_32_P190343	Unknown	Unknown
A_32_P227158	Unknown	Unknown
A_32_P874394	Unknown	Unknown
A_32_P30874	Unknown	Unknown

copula-based approach in finding genes that are correlated with quantitative outcome in melanoma studies. The main focus was on demonstrating how powerful the copula method is in selecting genes that are correlated with quantitative outcome while controlling for FDR. Simulations indicated that the copula-based model had reasonable power at various levels of the FDR. Our approach is flexible as no parametric assumption is made on the marginal distribution except that they are continuous. Relaxing parametric assumptions on microarray data may yield procedures that have good power for selecting differentially expressed genes. Although the methodology was motivated by data from the agilent technology, it can be adopted for data from any technology where both the gene expression levels and the outcome of interest are continuous.

New technologies such as RNA-sequencing (RNA-seq) are slowly replacing micoraary

Figure 4.2: **Venn diagrams of genes from different genelists** .(a) Intersection of copula G_2 25 genes and 165 from Omolo et al.,(2013) (b) Intersection of Cox genes, 4 copula list and 34 list from Omolo et al.,(2013).

technology. Current methods being developed for differential gene expression analysis are focusing on RNA-seq data. However, RNA-seq is more costly than microarrays (Bair, 2013). For a quick and easy experiment, microarrays can provide reliable and sensitive results. Therefore, new methods for analysing data from microarrays is till relevant and timely. Comparison of the proposed copula-based approach with the existing methods is done in the next chapter. We will also demonstrate how to choose an optimal copula for real microarray dataset.

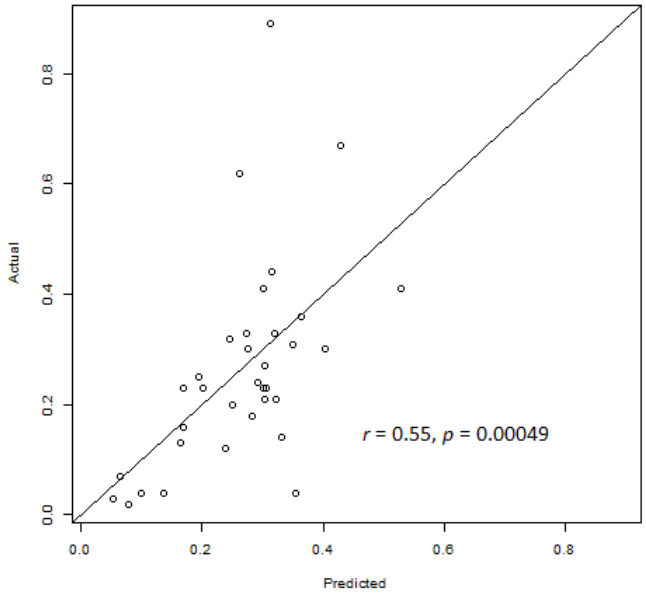

Figure 4.3: A scatter plot for actual values of G_2 values verses predicted values. Prediction of G_2 values were done using LASSO with LOOCV method.

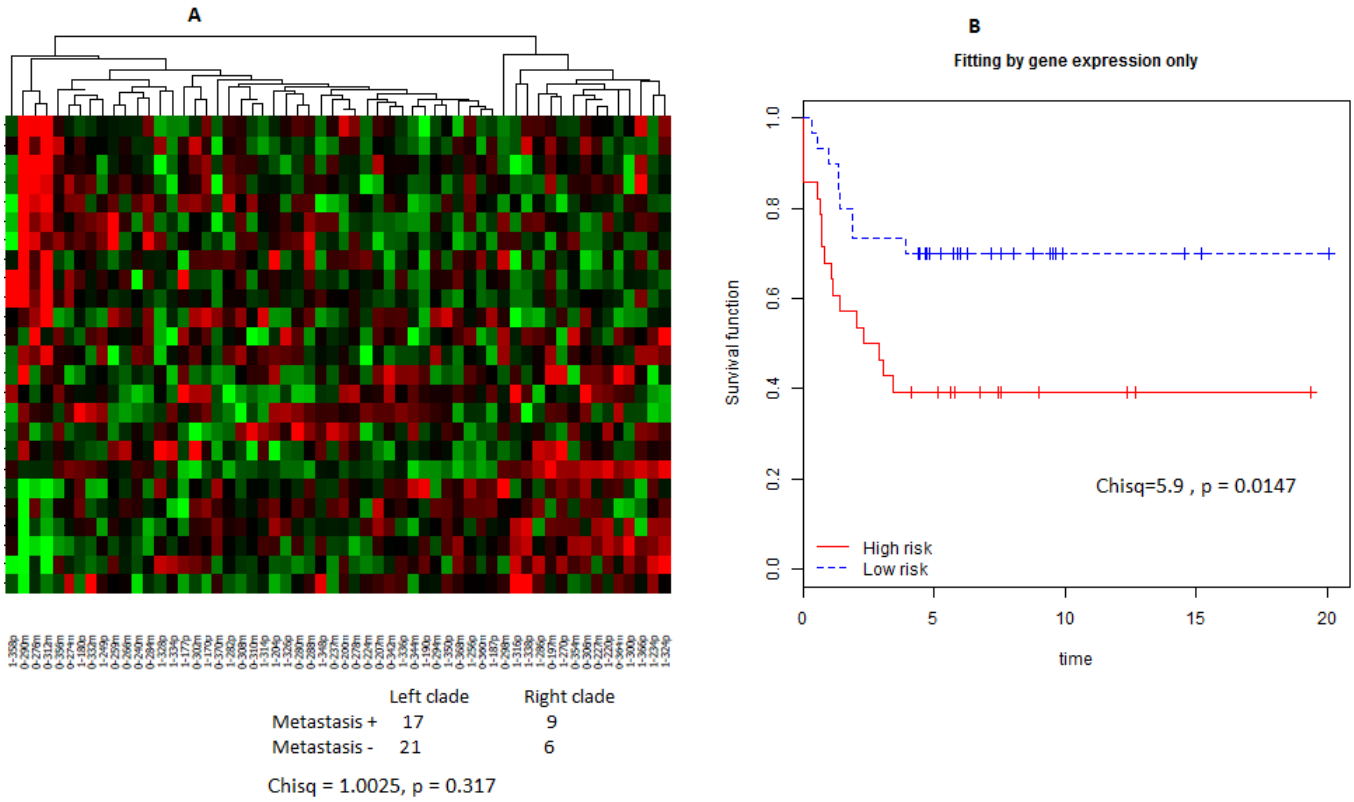

Figure 4.4: **Kaplan-Meier survival plot and heatmap for the copula gene signature.** (**A**): Unsupervised hierarchical cluster of 58 samples using the 25 copula genes that were correlated with G2 checkpoint function. The classification of the samples yielded non-significant results ($\chi^2 = 1.0025$, p = 0.317) . (**B**): The separation of the two groups in the Kaplan-Meier survival plot was significant ($\chi^2 = 5.9$, p = 0.0147).

Chapter 5

Comparison of the Copula Model with the QTA for Microarray Analysis

In this chapter we compare the copula-based approach for finding DEGs with the QTA method described in Chapter 2 by use of a simulation study. The QTA method was found to be a better method for finding genes that are good predictors of a quantitative outcome (Chaba et al., 2017). We use power and control of Type I error as the main basis of the comparison. These two measurements may indicate how the copula-based approach performs in comparison with the QTA method.

5.1 Which Copula to Use?

Having an appropriate copula in copula modelling is very crucial. To date, no study has been conducted on choosing the best copula model for gene expression data analysis. In the literature where the copula is applied on gene expression data, the choice of the copula is done arbitrarily. Some authors chose the copula based on how convenient they were for the analysis (Owzar et al., 2007). Others like Yuan et al. (2008) chose the copula based on the value of the likelihood. They chose the copula with the largest likelihood.

With several copulas to choose from in empirical applications, one needs an appropriate one for the data at hand. The statistical features of the data should guide the selection of the copulas. For example, gene expression profiles can be positively or negatively associated with a quantitative outcome. Therefore, naturally, copulas that can capture both the negative and the positive dependence such as the Normal copula, the Student-t copula and the Frank copula should be superior to the Gumbel and Clayton copulas, which do not permit negative dependence. To this end, we recommend the

following procedure:

1. Perform copula model selection based on the existing methods on all the pairs (a quantitative outcome and each gene expression profile). This helps in determining the closest parametric copula family from the list of copulas provided.

2. Record the proportion of pairs that are fitted by different parametric copulas.

3. The copula that fits most of the pairs is assumed for the whole analysis.

We consider two copulas: the Normal copula and the Frank copula, since they permit both positive and negative dependence. We perform model selection based on the AIC and the BIC, using the melanoma cell lines dataset. The Student-t copula is close to Normal copula, hence was not considered. The copula that fits the highest proportion of the pairs is adopted for the comparison of the two gene selection methods. The goodness-of-fit-test for the two copulas was also performed, using the Cramer-Von Mises (CVM) function Genest et al. (2009). Given the results in Table 5.1, the normal copula is the best of the two and will be used in the comparison.

Table 5.1: *Copula model selection based on three methods.*

Model selection method	Normal copula	Frank copula
CVM	35.18%	23.18%
AIC	58.40%	41.60%
BIC	58.40%	41.60%

5.2 Simulated Gene Expression Data

Let n and G denote the number of samples and genes, respectively. Further, let D denote the number of genes assumed to be truly differentially expressed. Then $(G - D)$ genes are assumed to be non-differentially expressed. The gene expression data matrix, $\mathbf{X}$, is a $G \times n$ matrix of log2-ratios. We can write $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$, where $\mathbf{X_1}$ and $\mathbf{X_2}$ are $D \times n$ and $(G - D) \times n$ matrices, respectively. We set $D \in (50, 100, 200, 300, 400)$, $n = 35$ and G to be 1000. We generated the $(1000 - D)$ genes from the standard normal distribution. To generate the D genes, we used the standard normal distribution in conjunction with the Cholesky decomposition (Golub and Van Loan, 1996) of their correlation matrix as follows:

1. We generate an unstructured correlation matrix $\mathbf{\Omega}$. $\mathbf{\Omega}$ is a $(D+1)\times(D+1)$ matrix that has $(i,j)^{th}$ element given by $\omega_{i,j} = \text{corr}(x_i, x_j)$
2. Find the Cholesky factor, $\mathbf{A}$, of $\mathbf{\Omega}$ such that $\mathbf{\Omega} = \mathbf{A}\mathbf{A}'$.
3. Let $\mathbf{z}_i \sim N(\underline{0}, \mathrm{I}_n), i = 1, 2, ..., (D+1)$.
4. $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{D+1})'$
5. $\mathbf{X}_{D+1} = \mathbf{A}\mathbf{Z}$. $\mathbf{X}_{D+1}$ is the gene expression matrix for D genes assumed to be differentially expressed and a covariate $\mathbf{y}$. $\mathbf{y}$ can take any of the $D+1$ row vectors from the matrix $\mathbf{X}_{D+1}$. $\mathbf{X}_1$ is therefore a submatrix of $\mathbf{X}_{D+1}$ with dimensions $D\times n$.

This simulation set-up assumes that each gene is observed across each sample. We compare our copula-based approach with the QTA method.

5.2.1 Simulation Results

Table 5.2 reports the number of genes declared to be differentially expressed for the copula and the QTA methods at different levels of FDR threshhold. The results indicate that, in general the copula method identified more DEGs than the QTA method. We note that the identified DEGs is likely to include both the truly DEGs and the false positives.

In order to evaluate the two methods in terms of power, the number of DEGs and the number of truly DEGs were recorded. The power was then calculated as the ratio of the number of truly DEGs to the total number of known DEGs. Power comparison results are shown in Table 5.3. Both the copula and the QTA methods have sufficient power to detect DEGs with a power of 1 in most cases. A power of 1 means that the method is able to detect all the known DEGs. In cases where the power was different for the two methods, the copula method stood out. This is seen especially when $D = 200$.

Controlling Type I error here means having an empirical Type I error close to the nominal level of the test. The closer the empirical Type I error is to the set nominal level, the better the method in controlling it. Table 5.4 shows that both the methods have reasonably controlled for Type I error at different nominal levels. There is evidence of both over and under estimation of the nominal levels by both methods although the the deviations are minimal. We note that the copula method consistently estimated the 0.01 nominal level for all the values of D except for $D = 50$ as compared to the QTA method. We also see that the accuracy of controlling the Type I error rate for the copula method increases with the increase in the number of known DEGs. This means that, even with a large number of DEGs identified by the copula method, we still trust the

copula approach to properly control the Type I error rate. The closeness of the results for the two methods on the simulated datasets may be due to the fact the data was generated from a normal distribution. Note that a bivariate Gaussian copula with two normal marginals corresponds to a bivariate Gaussian distribution. As such, the copula parameter reduces to the linear correlation coefficient (Pearson's correlation coefficient). The QTA method calculates its p-values based on either the Pearson's correlation coefficient or the Spearman's rho correlation coefficient.

Table 5.2: *Number of DEGs by the copula and the QTA methods at different estimated FDR levels.*

		Estimated FDR threshold			
D	Method	0.01	0.05	0.1	0.2
50	Copula	48	52	53	59
	QTA	50	53	57	63
100	Copula	100	106	110	126
	QTA	99	103	109	116
200	Copula	192	221	240	276
	QTA	146	204	221	241
300	Copula	302	314	326	378
	QTA	284	306	315	348
400	Copula	403	423	449	506
	QTA	405	415	429	466

5.3 Real Data Analysis

In this section, we use the datasets described in the Chapter 2. To find DEGs, we apply the same procedures for the QTA method and the copula method as described in Chapters 2 and 4 respectively. We note that the QTA method assumes normality on the marginal as well as linearity on the relationship between two variables. This is not always the case especially for gene expression levels data. In Figure 5.1, none of the randomly selected genes showed a linear relationship with the G_2 checkpoint function.

Table 5.5 shows the number of DEGs identified by the copula and the QTA methods based on the melanoma cell lines data and the G_2 checkpoint function as the quantitative outcome. In detecting DEGs using the real melanoma dataset, the copula method

Table 5.3: *Power comparison for the copula method and the QTA method for different nominal level and different number of known DEGs*

		Estimated FDR threshold			
D	Method	0.01	0.05	0.1	0.2
50	Copula	0.96	1.00	1.00	1.00
	QTA	1.00	1.00	1.00	1.00
100	Copula	0.99	1.00	1.00	1.00
	QTA	0.99	1.00	1.00	1.00
200	Copula	0.95	1.00	1.00	1.00
	QTA	0.72	0.97	1.00	1.00
300	Copula	1.00	1.00	1.00	1.00
	QTA	0.94	0.99	1.00	1.00
400	Copula	1.00	1.00	1.00	1.00
	QTA	1.00	1.00	1.00	1.00

Table 5.4: *Type I error rates of our method compared to the QTA method for different nominal level and different number of known DEGs*

		Estimated FDR threshold			
D	Method	0.01	0.05	0.1	0.2
50	Copula	0.00	0.04	0.06	0.15
	QTA	0.00	0.06	0.12	0.21
100	Copula	0.01	0.06	0.09	0.21
	QTA	0.00	0.03	0.08	0.14
200	Copula	0.01	0.10	0.17	0.28
	QTA	0.02	0.05	0.10	0.17
300	Copula	0.01	0.04	0.08	0.21
	QTA	0.00	0.03	0.05	0.14
400	Copula	0.01	0.05	0.11	0.21
	QTA	0.01	0.04	0.07	0.14

identifies more DEGs than the QTA especially at very low FDR thresholds. However as the FDR increases, QTA identifies more DEGs than the copula method. Figure 5.2 shows that an intesection of the genelists generated by the two methods yields 11 (15.7%)

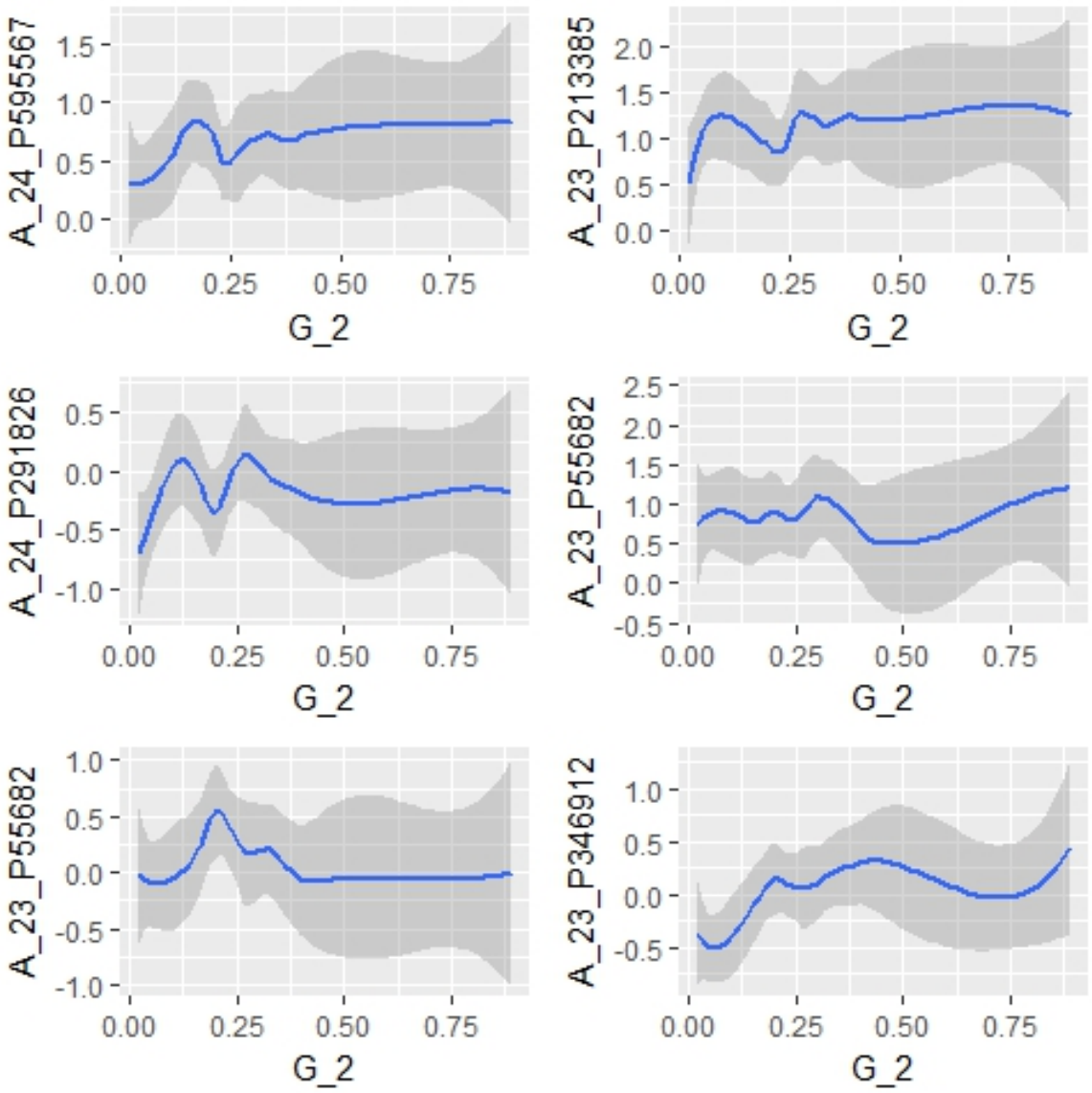

Figure 5.1: Expression profiles of a few genes as a function of the quantitative outcome (G_2). Gene expressions are associated with the G_2 in a nonlinear manner.

common genes. The results also indicate that the copula approach is a better predictor of a sample being in either low or high risk group of developing distant metastasis than the QTA method. QTA however shows better results in predicting G_2 checkpoint function than the copula method.

Table 5.5: *Number of genes declared differentially expressed using the copula method and the QTA method on the melanoma cell lines dataset, prediction results based on the LASSO with LOOCV and the survival risk prediction results based on log rank test are also provided*

	Estimated FDR threshhold				Prediction of G_2		SRP	
Method	0.01	0.05	0.1	0.2	r	p-value	Chis	p-value
Copula	9	9	9	25	0.5	0.0022	6.7	0.0096
QTA	0	0	4	56	0.721	<0.001	4.8	0.0374

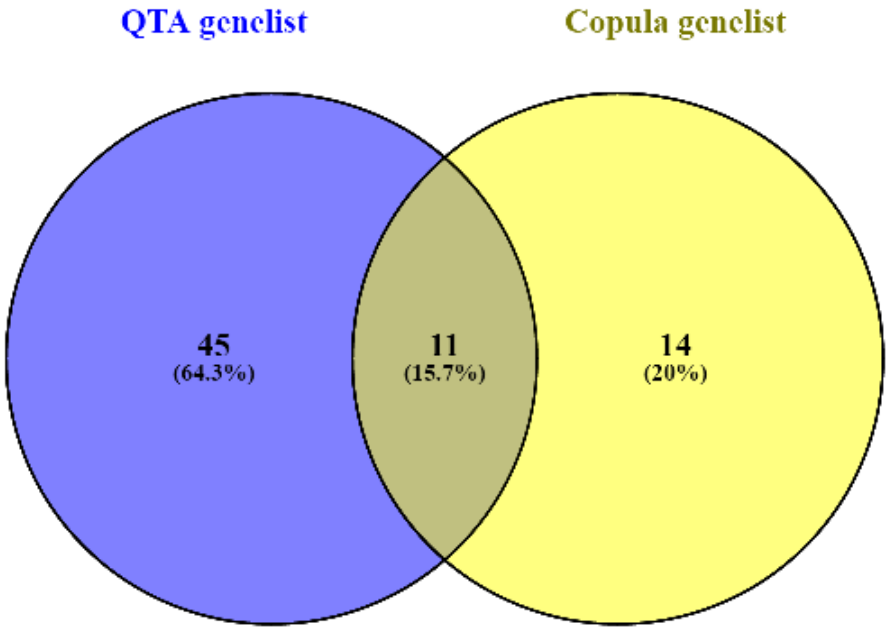

Figure 5.2: Number of overlapping genes from the copula and the QTA genelists based on the melanoma cell lines dataset.

5.4 Conclusion

This chapter presents a comparison of our proposed copula-based approach with the QTA method for finding differentially expressed genes when the outcome is continuous in nature. Using the power and the control of Type I error as a baseline for comparison, both methods perform well in power comparison but the copula approach was notably the better. In terms of the Type I error rate control, the two methods are comparable. We have also proposed a simple way of choosing a copula for gene expression studies. This approach is however limited to the copulas that permit both negative and positive depen-

dence only, and therefore better methods need to be developed. Both the copula method and the QTA are useful in generating gene signatures that are clinically important.

It is reasonable to conclude that, based on the current study, a semi-parametric copula approach outperform a QTA model which is parametric, in noisy high-dimensional data settings like in microarray studies. Here, however, we are limited to the QTA model, but other parametric models do exist (e.g. Bayesian models, etc.). It would therefore be interesting to see how semi-parametric models perform when compared to Bayesian models, in particular, in a future study.

Chapter 6

Summary

6.1 Main Findings

We have proposed a new algorithm for finding differentially expressed genes based on the copula functions. The main purpose was to develop a new procedure that is able to identify genes that are correlated with a quantitative outcome. The developed algorithm was evaluated using simulated datastes and applied on the real datasets for validation. We have shown that our proposed method has a good power to select DEGs and controlled the Type I error rate well for the settings we considered. We have also demonstrated the effectiveness of our copula-based approach in the analysis of a real data set. The genelist generated by our proposed approach is a good predictor of a quantitative outcome and is prognostic. The identification of such types of genes may lead to a more accurate diagnostics and treatment at individual patient level. In general, the simulation and data analysis results suggest that the proposed copula-based algorithm is a promising approach in differential gene expression analysis.

Prior to our proposal, we evaluated and compared four existing methods for finding genes when the outcome of interest is continuous. These four methods are the SAM, the LIMMA, the LPC and the QTA. The main reason for performing the evaluation was to find the "best" performing method to be compared with our proposed copula method. In this regard, the QTA approach was selected based on its predictive ability. We have shown that the algorithm proposed outperforms the QTA in terms of power, but the performance is comparable in terms of Type I error rate control. Compared with existing methods, our method is flexible in that one does not need to specify the marginal distribution as long as it is continuous. From all the comparative analysis in this study, it is clear that the choice of the gene selection method dictates the genes that are identified as differentially expressed.

6.2 Limitations

This is the first time this approach is applied to the analysis of gene expression data with a quantitative outcome. Therefore, issues are expected to arise and they will need to be addressed. The major downside of our algorithm is the high computational burden it bears as the size of the gene expression data gets large. This is because the gene specific p-values are calculated using permutation resampling method. The higher the number of permutations, the slower the process. The second drawback is the use of an assumed copula for the analysis of the whole dataset. The dependence structure between each gene expression level and any quantitative outcome is not fixed, but rather vary from one gene to the other. Therefore, using an assumed parametric copula to model the dependence of all pairs may not be appropriate even after choosing an optimal copula based on the steps provided in this work. We have also assumed that genes are independent. This is not always true for microaray studies as genes are believed to work together in a pathway.

6.3 Possible Extensions

Choosing the best copula to find genes that are correlated with quantitative trait is still an open field and more studies need to be conducted in this area. It would also be interesting to extend the copula-based approach to finding differentially expressed genes expression for RNA-sequencing (RNA-seq) dataset which is discrete in nature. This should be done with caution since there does not exist a unique copula identifying the joint distribution function of discrete variables.

Appendices

Appendix A

A bivariate Normal Copula

A bivariate normal copula is expressed as

$$C(u_1, u_2; \theta) = \Phi_\theta(\Phi^{-1}(u_1), \Phi^{-1}(u_2)), \tag{6.1}$$

where

$$\Phi_\theta = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left[-\frac{x^2 - 2\theta xy + y^2}{2(1-\theta^2)}\right] \mathrm{d}x\mathrm{d}y \tag{6.2}$$

is the bivariate standard normal distribution function with the correlation parameter $\theta \in [-1, 1]$ and

$$\Phi(u_i) = \int_{-\infty}^{\Phi^{-1}(u_i)} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right] \mathrm{d}x \; i = 1, 2 \tag{6.3}$$

denotes the univariate standard normal distribution function. We find probability density function of copula, $c(u_1, u_2; \theta)$ by differentiating the $C(u_1, u_2; \theta)$ with respect to u_1 and u_2. i.e.

$$c(u_1, u_2; \theta) = \frac{\partial C(u_1, u_2; \theta)}{\partial u_1 \partial u_2}. \tag{6.4}$$

Let $q_i = \Phi^{-1}(u_i)$. Therefore,

$$c(u_1, u_2; \theta) = \frac{\partial \Phi(q_1, q_2)}{\partial q_1 \partial q_2} \frac{\partial q_1}{\partial u_1} \frac{\partial q_2}{\partial u_2}. \tag{6.5}$$

But

$$\frac{\partial q_i}{\partial u_i} = \frac{\partial \Phi^{-1}(u_i)}{\partial u_i} = \left(\frac{\partial \Phi(q_i)}{\partial q_i}\right)^{-1} \tag{6.6}$$

Equation 6.5 therefore becomes

$$c(u_1, u_2; \theta) = \frac{\partial \Phi}{\partial q_1 \partial q_2} \left(\frac{\partial \Phi(q_1)}{\partial q_1} \right)^{-1} \left(\frac{\partial \Phi(q_2)}{\partial q_2} \right)^{-1}. \tag{6.7}$$

The copula density function thus becomes

$$c(u_1, u_2; \theta) = \frac{1}{\sqrt{1-\theta^2}} exp \left[\frac{1}{2} (\Phi^{-1}(u_1))^2 + (\Phi^{-1}(u_2))^2 - \frac{(\Phi^{-1}(u_1))^2 + 2\theta \Phi^{-1}(u_1)\Phi^{-1}(u_2) + (\Phi^{-1}(u_2))^2}{2(1-\theta^2)} \right] \tag{6.8}$$

and the likelihood function in terms of the normal copula is

$$f(x_1, x_2) = c(u_1, u_2; \theta) \prod_{i=1}^{2} f_i(x_i). \tag{6.9}$$

$f(x_1, x_2)$ reduces to a bivariate normal if $f_i(x_i)$ is normal. The log-likelihood function becomes

$$\ell_n(\theta) = \sum_{j=1}^{n} \log c(u_{1j}, u_{2j}; \theta) + \sum_{j=1}^{n} \sum_{i=1}^{2} \log(f_i(x_{ij})). \tag{6.10}$$

The dependence parameter θ is then estimated as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \ell_n(\theta). \tag{6.11}$$

The Kendall's τ and Spearman's ρ for the normal copula are given as

$$\tau = \frac{2}{\pi} arcsin(\theta) \tag{6.12}$$

and

$$\rho = \frac{6}{\pi} arcsin(\frac{\theta}{2}), \tag{6.13}$$

respectively. For proofs of (6.12) and (6.13), see McNeil et al. (2005).

Appendix B

Simulation R code

```
#For n=35 and D=50. The rest of the setting is carried out using the sam
approach
#########################################################################
x1 <- matrix(rnorm(35*950), nrow=950)
require(Matrix)


#Generating a correlation matrix:
#########################################################################
R <- matrix(runif(51*51), ncol=51)
RtR <- R %*% t(R)
Q <- cov2cor(RtR)#Correlation matrix
#########################################################################
#Cholesky decomposition
#########################################################################
U = t(chol(Q))
nvars = dim(U)[1]
numobs = 35
random.normal = matrix(rnorm(nvars*numobs,0,1), nrow=nvars, ncol=numobs);
x2 = U %*% random.normal
class(x2)
########################################
#Data with 1000 genes and 35 samples
########################################
```

```
Exp=rbind(x2[2:51,],x1[1:950,])#with genes on the rows and samples on the columns
plot_ly(z = Exp, type = "heatmap")#ploting heatmap
tExp=t(Exp)#with genes on the columns and samples on the rows
write.csv(tExp,file="tExp2.csv")
G22= x2[1,] #representing quantitative outcome
write.csv(G22,file="G22.csv")
```

Appendix C

Copula R code

```
rm(list=ls())
#Setting working directory
#########################################################
setwd("C:\\Users\\Linda\\Documents\\copula paper\\Melanoma dataset analysis")
dir()
 #Load copula library
#############################################################
library(copula)
#########################################################################
#loading data
#########################################################################################
g2=read.csv("g2.csv",row.names=1) y=g2[,1]# vector of quantitative outcomes
x=read.csv("NewFiltered35ArrayDataG2.csv",row.names=1) #A matrix of gene
expression data (3860 genes)

#########################################################################################
#function for calculating copula parameter. Use mpl method. Matrix chosen to
```

```
be exchangeable. #The dispersion matrix won't be a big deal even if its
unstructured since its just a 2 by 2 matrix
##############################################################################
cop.theta.i<-function(expr,QT)
 {
  u=pobs(cbind(expr,QT))
  cop2=normalCopula(dim=2,dispstr="ex")
  res=fitCopula(cop2, u, method = "mpl")
  coef(res)
}

cop.theta<-function(EXPR,QT)
 {
  apply(EXPR,2,cop.theta.i,QT)
}
theta=cop.theta(t(x)[,1:1000],y)

#Permuattion resampling
################################################
perm.index<-function(B,y)
{
  perm.index.i<-function(i,y)
  {
    sample(y,replace=FALSE)
  }
  sapply(1:B,perm.index.i,y)
}
```

```
yp=perm.index(1000,y)
############################################################
#setting up parallel computing
############################################################
library(doParallel)
cl <- makeCluster(3)
registerDoParallel(cl)
getDoParWorkers()
clusterExport(cl, list("y","x","cop.theta.i"))
clusterEvalQ(cl, library(copula))

#########################################################################
#theta from permuted data
#########################################################################
pthetag1=foreach(i=1:1000,.combine=cbind) %dopar%  cop.theta(t(x)[,1:1000],
yp[,i])

#Finding p values
#################################################################
rawpvalues=foreach(i=1:nrow(ptheta), .combine=rbind)%dopar%  mean(as.numeric
(ptheta[i,]>=thetanot[i]))

#Calculating q-values (estimated FDR)
#################################################################
source("https://bioconductor.org/biocLite.R") biocLite("qvalue")
library(qvalue) qobj=qvalue(rawpvalues, pi0.method="bootstrap")
```

```
summary(qobj)
OUT<-data.frame(qobj$pvalues,qobj$qvalues)
colnames(OUT)=c("pvalues","qvalues") #OUT11<-OUT11[order(qobj$qvalues),]
write.csv(OUT,"qvalueslpcsimulation2.csv")
```

Appendix D

SAM R code

```
rm(list=ls())

#Setting working directory
#########################################################################
setwd ("C:\\Users\\Linda\\Documents\\review paper\\Analysis\\SAM")
dir()
#Reading in the data
#########################################################################
x=read.csv("NewFiltered35ArrayDataG2.csv",row.names = 1)# Expression values
data with column one as the gene ID
x1=x[,1:35]#Getting just the expression data without the rownames
x1=as.matrix(x1)#Samr needs data in matrix format and not data frame
y=read.csv("g2.csv",row.names=1)#Bringing in outcome data. Excel file with
sample names, and two outcomes
attach(y)
y1=G2#Getting the outcome of interest G2

#Loading samr package
#########################################################################
```

```
library(samr)
#########################################################################
sam.data<-list(x=x1,y=y1, geneid= as.character(1:nrow(x)),
genenames= row.names(x))#Format needed by samr
samr.obj<-samr(sam.data, resp.type="Quantitative", regression.method="standard",
 knn.neighbors=10, nperms=1000, random.seed=1234567)
delta.table <- samr.compute.delta.table(samr.obj)
siggenes.table<-samr.compute.siggenes.table(samr.obj, 0.00, sam.data, delta.table)
res0.00 <- rbind(siggenes.table$genes.up,siggenes.table$genes.lo)
write.csv(res0.00,"res0.00.csv")
write.csv(delta.table,"deltatable0.00.csv")
################################################
```

Appendix E

LIMMA R code

```
rm(list=ls())
setwd ("C:\\Users\\Linda\\Documents\\review paper\\Analysis\\LIMMA")
dir()
#####################################################################
#Reading in the data
#####################################################################
x=read.csv("NewFiltered35ArrayDataG2.csv",row.names = 1)# Expression values
data with column one as the gene ID
x1=x[,1:35]#Getting just the expression data without the rownames
x1=as.matrix(x1)#Samr needs data in matrix format and not data frame
y=read.csv("g2.csv",row.names=1)#Bringing in outcome data. Excel file with
```

```
sample names, and two outcomes
attach(y)
y1=G2#Getting the outcome of interest DecatG2

#####################################################################
#Loading limma package
#####################################################################
source("https://bioconductor.org/biocLite.R")
biocLite("limma")
library(limma)
#####################################################################
eset=t(y1)#Transpose covariate, format required by limma
#MA is expression data only
MA=x[,1:35]
#Creating a design matrix. Limma needs specification of design matrix.
This requires spline package
#####################################################################
require(splines)
X <- ns(eset, df=5)
design <- model.matrix(~X)
#fitting the model
fit <- lmFit(MA, design)
fit <- eBayes(fit)
#Getting top significant genes based on fdr value of
results = topTable(fit, adjust= "fdr",n=Inf)
write.csv(results, file="limmalist.csv")
#####################################################################
```

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Andrew, H., G, F., and Golum, K. B. (2015). Methods for Identifying Differentially Expressed Genes: An Empirical Comparison. *Journal of Biometrics & Biostatistics*, 06(05).

Bair, E. (2013). Identification of significant features in DNA microarray data: Feature selection in DNA microarray data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(4):309–325.

Bair, E. and Tibshirani, R. (2004). Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. *PLoS Biology*, 2(4).

Baldi, P. and Long, A. D. (2001). A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509519.

Bandyopadhyay, S., Mallik, S., and Mukhopadhyay, A. (2014). A survey and comparative study of statistical tests for identifying differential expression from microarray data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 11(1):95–115.

Bao, L., Zhu, Z., and Ye, J. (2009). Modeling oncology gene pathways network with multiple genotypes and phenotypes via a copula method. In *Computational Intelligence in Bioinformatics and Computational Biology, 2009. CIBCB'09. IEEE Symposium on*, pages 237–246. IEEE.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Berg, D. (2009). Copula goodness-of-fit testing: an overview and power comparison. *The European Journal of Finance*, 15(7-8):675–701.

Bogunovic, D., O'Neill, D. W., Belitskaya-Levy, I., Vacic, V., Yu, Y.-L., Adams, S., Darvishian, F., Berman, R., Shapiro, R., Pavlick, A. C., and others (2009). Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. *Proceedings of the National Academy of Sciences*, 106(48):20429–20434.

Carson, C., Omolo, B., Chu, H., Zhou, Y., Sambade, M. J., Peters, E. C., Tompkins, P., Simpson, D. A., Thomas, N. E., Fan, C., Sarasin, A., Dessen, P., Shields, J. M., Ibrahim, J. G., and Kaufmann, W. K. (2012). A prognostic signature of defective p53-dependent G1 checkpoint function in melanoma cell lines: A signature of defective p53 function in melanoma. *Pigment Cell & Melanoma Research*, 25(4):514–526.

Chaba, L., Odhiambo, J., and Omolo, B. (2017). Evaluation of Methods for Gene Selection in Melanoma Cell Lines. *International Journal of Statistics in Medical Research*, 6(1):1–9.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula Methods in Finance*. The Wiley Finance Series. Wiley.

Conlon, E. M., Song, J. J., and Liu, A. (2007). Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics*, 8(1):80.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.*, 14(4):457–460.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412(6849):822–826.

Ding, C. H. (2003). Unsupervised feature selection via two-way ordering in gene expression analysis. *Bioinformatics*, 19(10):1259–1266.

Dobri, J. and Schmid, F. (2007). A goodness of fit test for copulas based on Rosenblatt's transformation. *Computational Statistics & Data Analysis*, 51(9):4633–4642.

Dubey (1994). Adjustment of p-values for multiplicities of intercorrelating symptoms. *Statistics in the Pharmaceutical Industry*, page 513527.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, 12(1):111–140.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):11511160.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868.

Emura, T. and Chen, Y.-H. (2016). Gene selection for survival data under dependent censoring: A copula-based approach. *Statistical Methods in Medical Research*, 25(6):2840–2857.

Escarela, G. and Carrire, J. F. (2003). Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research*, 12(4):333–349.

Fang, K., Kotz, S., and Ng, K. (1990). *Symmetric multivariate and related distributions*. Monographs on statistics and applied probability. Chapman and Hall.

Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis*, 95(1):119–152.

Genest, C., Ghoudi, K., and Rvest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Genest, C., Quessy, J.-F., and Remillard, B. (2006). Goodness-of-fit Procedures for Copula Models Based on the Probability Integral Transformation. *Scandinavian Journal of Statistics*, 33(2):337–366.

Genest, C., Rmillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2):199 – 213.

Golub, G. and Van Loan, C. (1996). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press.

Golub, S. (1999). Molecular classifcation of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 289:531–537.

HealthGrove (2017). Malignant skin melanoma in kenya. `http://global-disease-burden.healthgrove.com/l/36528/Malignant-Skin-Melanoma-in-Kenya`. [Online; accessed 2017-01-12].

Ibrahim, J. G., Chen, M.-H., and Gray, R. J. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, 97(457):88–99.

Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., and Guedj, M. (2010). Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. *PLoS ONE*, 5(9):e12336.

Jeffery, I. B., Higgins, D. G., and Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC bioinformatics*, 7(1):359.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419.

John, T., Black, M. A., Toro, T. T., Leader, D., Gedye, C. A., Davis, I. D., Guilford, P. J., and Cebon, J. S. (2008). Predicting Clinical Outcome through Molecular Profiling in Stage III Melanoma. *Clinical Cancer Research*, 14(16):5173–5180.

Jönsson, G., Busch, C., Knappskog, S., Geisler, J., Miletic, H., Ringnr, M., Lillehaug, J. R., Borg, k., and Lnning, P. E. (2010). Gene Expression ProfilingBased Identification of Molecular Subtypes in Stage IV Melanomas with Different Clinical Outcome. *Clinical Cancer Research*, 16(13):3356–3367.

Jung, S.-H., Owzar, K., and George, S. L. (2005). A multiple testing procedure to associate gene expression levels with survival. *Statistics in Medicine*, 24(20):3077–3088.

Kaufmann, W. K., Carson, C. C., Omolo, B., Filgo, A. J., Sambade, M. J., Simpson, D. A., Shields, J. M., Ibrahim, J. G., and Thomas, N. E. (2014). Mechanisms of chromosomal instability in melanoma: Chromosomal Instability in Melanoma. *Environmental and Molecular Mutagenesis*, 55(6):457–471.

Kaufmann, W. K., Nevis, K. R., Qu, P., Ibrahim, J. G., Zhou, T., Zhou, Y., Simpson, D. A., Helms-Deaton, J., Cordeiro-Stone, M., Moore, D. T., Thomas, N. E., Hao, H., Liu, Z., Shields, J. M., Scott, G. A., and Sharpless, N. E. (2008). Defective Cell Cycle Checkpoint Functions in Melanoma Are Associated with Altered Patterns of Gene Expression. *Journal of Investigative Dermatology*, 128(1):175–187.

Kendziorski, C. M., Newton, M. A., Lan, H., and Gould, M. N. (2003). On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in medicine*, 22(24):3899–3914.

Kim, J.-M., Jung, Y.-S., Sungur, E. A., Han, K.-H., Park, C., and Sohn, I. (2008). A copula method for modeling directional dependence of genes. *BMC Bioinformatics*, 9(1):225.

Kim, S. Y., Lee, J. W., and Sohn, I. S. (2006). Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. *Statistical Methods in Medical Research*, 15(1):3–20.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398.

Le, C. T., Pan, W., and Lin, J. (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & Integrative Genomics*, 3(3):117–124.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003). Gene selection: a bayesian variable selection approach. *Bioinformatics*, 19(1):90–97.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98(1):31–36.

Li, M., Boehnke, M., Abecasis, G. R., and Song, P. X.-K. (2006). Quantitative Trait Linkage Analysis Using Gaussian Copulas. *Genetics*, 173(4):2317–2327.

Louie, H. (2014). Evaluation of bivariate Archimedean and elliptical copulas to model wind power dependency structures. *Wind Energy*, 17(2):225–240.

Mandruzzato, S., Callegaro, A., Turcatel, G., Francescato, S., Montesco, M. C., Chiarion-Sileni, V., Mocellin, S., Rossi, C. R., Bicciato, S., Wang, E., Marincola, F. M., and Zanovello, P. (2006). A gene expression signature associated with survival in metastatic melanoma. *Journal of Translational Medicine*, 4:50.

McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative risk management: concepts, techniques and tools*. Princeton series in finance. Princeton University Press, Princeton, N.J. OCLC: ocm60796246.

Nadon, R. and Shoemaker, J. (2002). Statistical issues with microarrays: processing and analysis. *TRENDS in Genetics*, 18(5):265–271.

Nelsen, R. B. (2006). *An Introduction to copulas*. Springer.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., and Blattner, F. R. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8:37–52.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176.

Omolo, B., Carson, C., Chu, H., Zhou, Y., Simpson, D. A., Hesse, J. E., Paules, R. S., Nyhan, K. C., Ibrahim, J. G., and Kaufmann, W. K. (2013). A prognostic signature of g_2 checkpoint function in melanoma cell lines. *Cell Cycle*, 12(7):1071–1082.

Owzar, K., Jung, S.-H., and Sen, P. K. (2007). A copula approach for detecting prognostic genes associated with survival outcome in microarray studies. *Biometrics*, 63(4):1089–1098.

Pan, W. (2003). On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics*, 19(11):1333–1340.

Qin, L.-X. and Kerr, K. F. (2004). Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Research*, 32(18):5471–5479.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47.

Romano, C. (2002). Applying copula function to risk management. In *Capitalia, Italy. http://www. icer. it/workshop/Romano. pdf.*

Scharpf, R. B., Tjelmeland, H., Parmigiani, G., and Nobel, A. B. (2009). A bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association*, 104(488):1295–1310.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A*, 93(20):10614–10619.

Schwartzman, A., Dougherty, R. F., and Taylor, J. E. (2008). False discovery rate analysis of brain diffusion direction maps. *The Annals of Applied Statistics*, 2(1):153–175.

Schwartzman, A. and Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika*, 98(1):199–214.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Schwender, H., Krause, A., and Ickstadt, K. (2003). Comparison of the empirical bayes and the significance analysis of microarrays. Technical report, Technical Report//Universitt Dortmund, SFB 475 Komplexittsreduktion in Multivariaten Datenstrukturen.

Sebastiani, P., Gussoni, E., Kohane, I. S., and Ramoni, M. F. (2003). Statistical challenges in functional genomics. *Statistical Science*, pages 33–60.

Serfling, R. J. (2002). *Approximation theorems of mathematical statistics*. Wiley series in probability and statistics. Wiley, New York, NY, paperback ed edition. OCLC: 49527610.

Shao, J. (2003). *Mathematical statistics*. Springer texts in statistics. Springer, New York, 2nd ed edition.

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer Statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67:7–30.

Simon, R., Lam, A., Li, M.-C., Ngan, M., Menenzes, S., and Zhao, Y. (2007). Analysis of Gene Expression Data Using BRB-Array Tools. *Cancer Informatics*, 3:11–17.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):Article 3.

Smyth, G. K. (2005). limma: Linear Models for Microarray Data. In Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A., and Dudoit, S., editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer New York, New York, NY.

Sreekumar, J.and Jose, K. K. (2008). Statistical tests for identification of differentially expressed genes in cDNA microarray experiments. *Indian Journal of Biotechnology*, 7(4):423–436.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Tarpey, P. S., Smith, R., Pleasance, E., Whibley, A., Edkins, S., Hardy, C., O'Meara, S., Latimer, C., Dicks, E., Menzies, A., Stephens, P., Blow, M., Greenman, C., Xue, Y., Tyler-Smith, C., Thompson, D., Gray, K., Andrews, J., Barthorpe, S., Buck, G., Cole, J., Dunmore, R., Jones, D., Maddison, M., Mironenko, T., Turner, R., Turrell, K., Varian, J., West, S., Widaa, S., Wray, P., Teague, J., Butler, A., Jenkinson, A., Jia, M., Richardson, D., Shepherd, R., Wooster, R., Tejada, M. I., Martinez, F., Carvill, G., Goliath, R., de Brouwer, A. P. M., van Bokhoven, H., Van Esch, H., Chelly, J., Raynaud, M., Ropers, H.-H., Abidi, F. E., Srivastava, A. K., Cox, J., Luo, Y., Mallya, U., Moon, J., Parnau, J., Mohammed, S., Tolmie, J. L., Shoubridge, C., Corbett, M., Gardner, A., Haan, E., Rujirabanjerd, S., Shaw, M., Vandeleur, L., Fullston, T., Easton, D. F., Boyle, J., Partington, M., Hackett, A., Field, M., Skinner, C., Stevenson, R. E., Bobrow, M., Turner, G., Schwartz, C. E., Gecz, J., Raymond, F. L., Futreal, P. A., and Stratton, M. R. (2009). A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nature Genetics*, 41(5):535–543.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. 1996;58(1):267288. *J Roy Statist Soc B.*, 58(1):267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, pages 104–117.

Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., and Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–1461.

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res*, 40(9):3785–3799.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

Wang, A. (2010). Goodness-of-fit tests for archimedean copula models. *Statistica Sinica*, 20(1):441.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Wigle, D. A., Jurisica, I., Radulovich, N., Pintilie, M., Rossant, J., Liu, N., Lu, C., Woodgett, J., Seiden, I., and Johnston, M. (2002). Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research*, 62(11):3005–3008.

Winnepenninckx, V., Lazar, V., Michiels, S., Dessen, P., Stas, M., Alonso, S. R., Avril, M.-F., Ortiz Romero, P. L., Robert, T., Balacescu, O., Eggermont, A. M. M., Lenoir, G., Sarasin, A., Tursz, T., van den Oord, J. J., and Spatz, A. (2006). Gene Expression Profiling of Primary Cutaneous Melanoma and Clinical Outcome. *JNCI Journal of the National Cancer Institute*, 98(7):472–482.

Witten, D. M. and Tibshirani, R. (2008). Testing significance of features by lassoed principal components. *The Annals of Applied Statistics*, 2(3):986–1012.

Xu, J. J. (1996). *Statistical Modelling and Inference for Multivariate and Longitudinal Discrete Response Data.* Ph.D Thesis, Department of Statistics, University of British Columbia,.

Yan, J. (2007). Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software*, 21(4):1–21.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(12):171–196.

Yuan, A., Chen, G., Zhou, Z.-C., Bonney, G., and Rotimi, C. (2008). Gene Copy Number Analysis for Family Data Using Semiparametric Copula Model. *Bioinformatics and Biology Insights*, 2:343–355.